THIS WEEK

EDITORIALS

PUBLISHING New online journal intends to promote data science **p.534**

WORLD VIEW The continuing saga of China and the environment **p.535**

BURIAL How Neanderthal funeral claim could be grave error **p.536**

A three-step plan for antibiotics

If the threat of antibiotic resistance is to be managed, existing drugs must be marshalled more effectively and new medicines must get to market fast.

If the first step towards solving a problem is to acknowledge its existence, then some important progress has been made on the thorny issue of antibiotic resistance. Last July, *Nature* noted approvingly a "notable rise in awareness among policy-makers and the public" on the issue and credited the advocacy of scientists for the surge (see *Nature* **499,** 379; 2013). That rise has continued, but with increased public and political awareness comes a greater demand for action. Much of that heavy expectation will fall on scientists. So, after the advocacy, how can the antibiotic-resistance threat be countered?

The first step, and one that must be pursued with urgency, is better stewardship of existing antibiotics. This demands fresh research and discoveries, but significant gains are also possible if officials and policy-makers can crank up the funds and willpower to match their rhetoric. Doctors and others who routinely overprescribe antibiotics for everything from sore throats to bronchitis need clear and explicit instructions from the top to stop. Medical schools that do not drum into their trainees the importance of prudence must start to do so.

It is not enough for doctors to urge their patients to finish the prescribed course when they are dishing out the pills with such abandon. A study published in the *Journal of the American Medical Association* last week showed that despite guidelines that veto such use of antibiotics for acute bronchitis and decades of research showing no benefit, the number of antibiotic prescriptions for this indication rose in the United States from 1996 to 2010 (M. L. Barnett & J. A. Linder *J. Am. Med. Assoc.* **311,** 2020–2022; 2014).

Over-the-counter sales of antibiotics must be banned. The countries that allow it are squandering a precious resource as surely as if they were tipping oil down the toilet. Regional regulations that limit the use of antibiotics to speed up the growth of livestock should extend worldwide. Public education — both to restrict the waste of antibiotics and to build support for measures to restrict unnecessary use — is vital. These are low-hanging fruit and they must be picked with all possible urgency. They need top-down political action, and that means governments. Cross-party consensus should be explicitly hammered out and publicized — there is no equivalent of Big Oil or Big Tobacco in this debate to delay and obfuscate.

The second step, and this is the one in which scientists have the biggest role, is to find ways to maximize the impact of our existing stocks. Researchers in the public and private sectors must re-examine all compound libraries for drugs that could couple with rapid diagnostic tests to offer new, narrow-spectrum therapies. Other compounds could be used in combination to reverse resistance to existing medicines and so extend their useful life — similar cocktails of drugs have been successful in treating HIV, after all.

Research can improve diagnosis too, to both speed up treatment of patients and minimize the waste of ineffective drugs. As a Comment article on page 557 points out, genome sequencing of infectious bacteria can rapidly identify resistance genes. So samples from an infected patient — analysed in clinical microbiology labs as close to point of care as possible — could steer drug treatment, at least in the developed world. Mass spectrometry was introduced for clinical use in this way a few years ago, it notes, and is now commonly used to identify pathogens from signature microbial peptides. Such a rapid front-line diagnostic kit to improve antibiotic use is one of the six major challenges identified by the UK government in its new Longitude Prize, intended to boost innovation.

The third step must be to boost the number of antibiotic drugs that are reaching the market. Between 1983 and 1992, the US Food and Drug Administration approved 30 new antibiotics; between 2003 to 2012, it approved just seven.

"Much of the heavy expectation will fall on scientists."

Reversing this trend is less about research and more about restructuring the financial incentives for firms to do that work. In 2012, for instance, the United States passed the Generating Antibiotic Incentives Now Act, which gives companies an extra five years of exclusive use for new antibiotics that they develop.

Others, including the World Health Organization (WHO), are considering more radical changes to the drug-development model itself. Last week, WHO members met to discuss a draft global action plan on antimicrobial resistance that floated "new business models" driven by public need rather than market forces. Such action would demand global consensus on the problem, and a Comment piece on page 555 argues that the globe needs a new body to help to achieve that and to drive action — an intergovernmental panel on antimicrobial resistance.

We have come a long way in a year. But the real work starts now. ■

Clean break

Improved biomass stoves are not popular, people everywhere deserve modern cooking methods.

For the billions of people who rely on food cooked over smoky open fires, a less-polluting stove seems like a clear solution. The devices allow people who have limited resources to use the same fuels — wood, charcoal, animal dung and agricultural waste — but generate less toxic fumes and therefore save millions of lives.

For decades, that apparent win–win strategy has held great appeal for big international donors, non-governmental organizations and engineers. This week, for example, the US Environmental Protection Agency announced grants to six universities for more research into clean-cooking stoves.

Unfortunately, these efforts are failing, at least on the broad scale.

Even though high-profile programmes have distributed millions of stoves to households in south Asia, Africa and Latin America, it is hard to find signs that the stoves are being widely used. There is a vast gap between reported accomplishments and what researchers see when they step into people's homes.

The crux of the problem is that simply supplying the stoves does not establish demand for them.

As a News Feature reports on page 548, women often complain that the stoves do not meet their needs. Some designs require wood to be chopped up into small pieces, thereby creating extra work; others do not burn hot enough, break easily or are too small or too expensive. Cooks from Bolivia to Bangladesh will use the stoves only if the devices make their lives easier. Too often, this is not the case, so the stoves get set aside — or are modified to work more like the traditional, pollution-producing stoves.

The downbeat assessment will not be popular with those who distribute the devices, such as the Global Alliance for Clean Cookstoves, a coalition based in Washington DC. But it should not come as a big surprise. In 2012, a report by the Massachusetts Institute of Technology in Cambridge, called *Up in Smoke*, found no long-term improvement in pulmonary health or in fuel savings among villagers who had received the stoves, mainly because people had abandoned the devices.

The alliance countered that the stoves just need to be adapted to meet local needs and that users need more training. The perpetual claim is that the biomass stove of people's dreams is just around the corner.

But some researchers looking at the health effects of cooking fires say that it is time for a fundamental shift in strategy — one that moves people away from burning biomass entirely.

Efforts could be redirected to providing people with the energy they most aspire to: not a stove designed by someone in the developed world to cook cleaner, but the actual stoves used in the developed world, which run on electricity or hydrocarbons such as liquefied petroleum gas (LPG).

"It is time for a fundamental shift in strategy — one that moves people away from burning biomass entirely."

This is not an absurd goal. The International Energy Agency (IEA) estimates that bringing electricity and clean-cooking facilities to every person on Earth by 2030 will cost US$49 billion a year. Although that is a considerable sum, the agency points to major commitments by Indonesia, Ghana and Nigeria to aggressively switch large portions of their population to cooking with LPG.

Where will all this new energy come from? It will require some additional consumption of fossil fuels, and that will increase the emissions of carbon dioxide into the atmosphere. But the extra pollution would be minimal at the global scale: the IEA estimates that it would boost CO_2 emissions by just 0.7% above its base scenario.

Renewable sources should be able to supply a major fraction of the needed energy: electrical micro-grids that use agricultural waste, solar cells or wind turbines to provide energy are popping up, for instance. Clean-cooking programmes have an enduring appeal, just not for their intended users. It is time to rethink the approach. ■

ANNOUNCEMENT

Welcome, *Scientific Data*!

Everybody is talking about data. Experimental scientists live and breathe data. Theorists are challenged by data. Funders are wondering how to make the data produced with their support more accessible without stretching their budgets. Research communities are seeking new data repositories, and standards to support them. And scientific publishers are wondering how to host data and provide quality control.

Scientific Data is a new journal, launched by *Nature*'s publishers this week, that will help to address some of these challenges. By publishing formal descriptions of data sets — Data Descriptors, the publication's main article type — it will render the data more visible and give originators explicit credit for those data, rather than for the papers that use them. The journal is peer-reviewed and online-only. Authors pay a charge on publication: this ensures that the final, published versions of their contributions to the journal are immediately freely accessible to all. The content is licensed under one of three Creative Commons licences, and machine-readable metadata are released with every article to maximize reuse.

To quote *Scientific Data*'s launch editorial: "The question is no longer whether research data should be shared, but how to make effective data sharing a common and well-rewarded part of research culture". When it is feasible to do so, many journals, including all those in the *Nature* family, have long insisted that data are deposited in repositories where available, before publication. For other areas of research, we at *Nature* have significantly increased the figure limits in our papers. In *Nature Protocols*, there is a place for more-specific methods descriptions than is conventional in scientific papers.

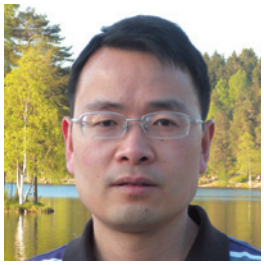
Now, in *Scientific Data*, there is space for researchers to formally describe a data set and the techniques used to derive it, and to refer readers to research papers that have already incorporated the data.

Crucially, the journal's descriptors, being peer-reviewed and citable, provide a way to assign credit to the originators of reusable data sets. In other words, the delivery and sharing of data becomes as credit-worthy, in principle, as publishing conventional research papers. It is important that the assessment of research and reward of researchers does more justice to this essential component of science.

The journal's first publications include articles describing previously unpublished data sets — demonstrations that *Scientific Data* can help to motivate scientists to share valuable data. The journal's editors highlight work by Zengchao Hao and colleagues detailing data sets that track drought around the world (Z. Hao *et al. Sci. Data* http://doi.org/sww; 2014). Using the Data Descriptor, anyone can download the data, generate their own maps (past or future) for any area of the world and even use the authors' source code to recalculate the drought metrics.

Another article, by Graham Edgar and Rick Stuart-Smith, provides an example of a Data Descriptor that builds on previous publications (G. J. Edgar and R. D. Stuart-Smith *Sci. Data* http://doi.org/sxv; 2014). It is based around the data produced by the Reef Life Survey, a citizen-science project that uses volunteer divers to help to survey biodiversity on the world's reefs. Analyses of these data, which are relevant to our understanding of reef ecology and to conservation, have been published in a number of research papers. The data are given in full in the Data Descriptor, along with the authors' descriptions of the survey procedures and data standardization — crucial information for other scientists interested in using these data.

Beyond its significance for data buffs, the journal is a further step in Nature Publishing Group's drive to enhance research reproducibility. The more researchers take steps to make their data available and discoverable, the more a core principle of science — that others can replicate the work — can be fulfilled, in an era in which such replication is often beset by obstacles. For that reason alone, we at *Nature* welcome *Scientific Data*. ■


XUEYAN HU

China must continue the momentum of green law

A plan for improved environmental protection is a good first step, but all levels of society will need to work together for it to succeed, says **Hong Yang**.

It has taken 25 years, but China is finally modernizing how it intends to protect the environment. Late last month, the country's legislature revised its Environmental Protection Law (EPL), potentially heralding the end of China's 'growth-at-any-cost' strategy.

As most people know, the cost of this strategy has been great. In 2012, China emitted more than one-quarter of the world's carbon dioxide. In April, a report from the Ministry of Environmental Protection and the Ministry of Land and Resources claimed that around 16% of the country's land and 20% of its farmland is contaminated with heavy metals and pesticides. Another report said that nearly 60% of monitored underground water was of "very poor" or "relatively poor" quality in 2013. Furthermore, smog routinely blankets Beijing and other cities. The pollution is not limited to China — it also disperses to Northern America and affects the weather across the Northern Hemisphere.

The revised legislation looks strong on paper. It introduces strict penalties and public shaming for polluters, with 15 days in jail for those in charge of companies found to be in breach of targets. Caps on fines, which made it cheaper for many companies to face prosecution than to install equipment to reduce pollution, have been scrapped. And some green groups and other non-governmental organizations (NGOs) have now been given the right to sue.

At the local level, government officials face greater scrutiny and must make public more information, such as environmental impact assessments done on proposed developments. And individuals are urged to do their bit, too: the provisions call on the public to recycle more of its rubbish, for instance.

As many commentators have pointed out, the revised law could help to steer China towards a more sustainable development path, but only if it is properly implemented and strictly enforced. Existing environmental standards have been crippled by a lack of political will and by corruption in local governments. Enforcement is likely to depend on robust action by regional bureaus, which are more exposed than central government to the lobbying and bribes of polluting industry.

The revised law may be accompanied by a restructuring of the government departments that have responsibility for the environment, which until now was overseen by several agencies, including the Ministry of Housing and Urban-Rural Development and the Ministry of Agriculture.

More effective than the restructuring could be wider use of environmental litigation, the possibilities for which remain limited under the updated rules. Only organizations registered above the city level will be able to launch lawsuits, so only around 300 NGOs qualify. Lawsuits are allowed only against polluters; the provisions for actions against enforcement authorities remain murky. A lawsuit brought against an environmental protection bureau for failing to curb air pollution was rejected in Shijiangzhuang, one of China's ten most polluted cities.

There is also a growing divide between urban and rural areas. Protests from the urban middle class have forced the suspension or relocation of many dirty factories, and an urbanization plan released in March will accelerate such relocation. Beijing, for instance, has ordered more than 50 heavy-machinery and chemical companies to move to less developed regions. The transfers are damaging the fragile middle and western parts of the nation: industrial wastewater discharge has doubled in Qinghai Province.

In April, the International Comparison Program for the World Bank projected that China was on route to become the world's biggest economy in terms of purchasing power by the end of the year. But economic success cannot hide the loss to the environment and to health. The 2004 *China Green National Accounting Study Report* put the cost of environmental degradation at 512 billion renminbi (US$62 billion; 3.1% of the total gross domestic product (GDP)). That tripled to about 1,539 billion renminbi in 2010, 3.5% of GDP. But because some audits are missing, for example on health loss, the actual costs could be much bigger. According to the United Nations Human Development Index, which includes the environment and health, China still performs poorly.

In announcing the revisions, the Chinese government promised a "war on pollution". But wars are expensive, and to keep that promise, the government will need to increase its investment: its spending on environmental protection and energy conservation fell by almost 10% between 2012 and 2013.

If the provisions are to have teeth, this trend will need to be reversed. Increased media attention and public pressure can help to achieve this, and scientists in China have an important role. More should speak out about the environmental crisis, both directly to the public and through the media. This could also help to avoid the perception in China that scientists are mouthpieces of the government.

The Chinese people are ready for a grown-up debate about environmental problems and the cost of fixing them. Vivid and widespread debate helped to shape China's revisions to its environmental legislation. This must be built on and extended to become a template for other laws, particularly those that seek to protect the air, soil and water. ■

THE REVISED LAW COULD HELP TO STEER CHINA TOWARDS A MORE SUSTAINABLE DEVELOPMENT PATH.

➲ **NATURE.COM**
Discuss this article online at:
go.nature.com/21biom

Hong Yang *is a scientist in the Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Norway.*
e-mail: hongyanghy@gmail.com

RESEARCH HIGHLIGHTS

Selections from the scientific literature

MATERIALS

Bent crystal gets back into shape

Scientists in Japan have discovered an organic crystal that can regain its structure after being deformed — the first known organic superelastic material.

Until now, superelastic materials, which change their crystal structure when mechanically stressed, were made only of metallic alloys and ceramics. Satoshi Takamizawa and Yasuhiro Miyamoto at Yokohama City University found that a terephthalamide crystal can also be superelastic.

The duo twisted the crystal, bending it and changing its molecular arrangement. When tension was released, it regained its structure without signs of material fatigue, even when the stress was applied and removed 100 times.

The advance could lead to self-repairing vehicle parts and materials that help to dampen vibrations, the authors say.

Angew. Chem. **http://doi.org/f2rqpt (2014)**

BEN THUY

PALAEONTOLOGY

Deep ocean is a safe haven

Deep-sea species might be more resilient to extinction than their shallow-water cousins.

Ben Thuy of the Luxembourg Natural History Museum and his team found fossils of at least 68 different species of molluscs, brachiopods, crustaceans and echinoderms (including relatives of a modern brittle star, *Ophiomyces frutectosus*, pictured) that lived more than 1,000 metres underwater around 190 million years ago. By comparing the fossils with shallow-living species from the same period, the team found that the deeper species tended to stay and diversify in deep waters, whereas shallow species were more likely to move to greater depths.

Some of the deep-sea species examined are the oldest known representatives of their respective families, implying that they might have originated and evolved in deep water, rather than migrating there from shallower seas as previously thought. The deep ocean could be a stable refuge for marine species, the team says.

Proc. R. Soc. B. 281, **20132624 (2014)**

ARCHAEOLOGY

Did Neanderthals bury their dead?

The remains of a Neanderthal in France may not have been buried ceremonially, as archaeologists had suggested.

A previous analysis of a nearly complete Neanderthal skeleton, found in a cave at La Chapelle-aux-Saints, concluded that the burial was intentional, noting that the depression looked dug-out and that the remains were well preserved.

JOHN READER/SPL

Harold Dibble at the University of Pennsylvania in Philadelphia and his colleagues now question this evidence. They found that the hole containing the skeleton (skull **pictured**) is much larger than would have been needed to hold a body. The hollow is also similar to brown-bear hibernation nests and to a second, smaller depression in the cave that holds bison remains. The researchers say that natural differences in weathering may explain why the Neanderthal remains in the cave are better preserved than those of other animals.

J. Archaeol. Sci. **http://doi.org/svx (2014)**

NANOTECHNOLOGY

Nanotubes form a complex circuit

Physicists have devised a way to use carbon nanotubes to build circuits that are more complex than previous attempts.

Highly conductive carbon nanotubes are a promising material to replace silicon transistors in integrated circuits. But when grown on a chip, not all nanotubes are semiconducting and their properties vary too much to make high-quality circuits. To get around this, Lian-Mao Peng, Zhiyong Zhang and their colleagues at Peking University in China made circuits out of smaller units, each made up of four pairs of transistors built on two nanotubes with different properties. The authors used these modules to build an 8-bit data-transfer system,

containing 46 transistors on six different nanotubes.

Although the method would be difficult to scale up with current materials, the authors say that the units could be used to explore the potential limits of integrated circuits based on carbon nanotubes.

Nano Lett. **http://doi.org/svv (2014)**

METABOLISM

A longer life with less pain

Mice that lack a class of pain receptors have a longer lifespan and a more youthful metabolism than normal mice.

Previous work had shown that, compared with control animals, mice missing TRPV1 pain receptors were less sensitive to pain and were less likely to become obese when fed a calorie-rich diet. Andrew Dillin of the University of California, Berkeley, and his colleagues have now found that such mice also live 12–16% longer than control mice.

The mutant mice make less of a molecule called CGRP in the pancreas. This, in turn, boosts insulin secretion, improving the control of blood-sugar levels. Furthermore, treating aged mice with an inhibitor of CGRP made their metabolic profiles look more like those of younger animals.

Cell 157, **1023–1036 (2014)**

ECOLOGY

Dancing bees reveal better land

Researchers have decoded the honeybee's dance to determine which types of land the insects prefer.

Honeybees do a 'waggle dance' (**pictured**) to tell their nestmates the best places to forage for nectar and pollen. Margaret Couvillon and her colleagues at the University of Sussex in Brighton, UK, observed foraging honeybees (*Apis mellifera*) from three colonies for two years. By analysing more than 5,000 waggle dances, the team found that the bees preferred tracts of land with greater stewardship, and a nature reserve with abundant wildflowers, to land sown with organic seed mixes and frequently mowed.

The authors suggest that bees could be used as indicators for improving environmental management.

Curr. Biol. **http://doi.org/sv9 (2014)**

REDMOND DURRELL/ALAMY

CLIMATE CHANGE

Soot drives Greenland melting

Forest fires in the Northern Hemisphere have helped to spur record surface melting of ice in Greenland.

Kaitlin Keegan of Dartmouth College in Hanover, New Hampshire, and her colleagues analysed six shallow ice cores from different areas of Greenland's interior. They found that both the record melting in 2012 and the last similarly widespread melting event in 1889 were caused by a combination of unusually warm air temperatures and increased heat absorption by snow laden with soot from distant forest fires.

Average summer temperatures and the frequency of forest fires in the Northern Hemisphere are projected to rise — and so, by the end of the century, widespread surface melting in Greenland might happen in almost any year, the authors say.

Proc. Natl Acad. Sci. USA **http://doi.org/svz (2014)**

ANIMAL BEHAVIOUR

Wild mice run for fun on wheels

Exercise wheels are popular among laboratory mice, but it seems that wild mice — and even frogs and shrews — also use them to go for a jog.

Some have argued that wheel-running is not a natural behaviour and could even be a result of neurosis. Johanna Meijer and Yuri Robbers at Leiden University Medical Centre in the Netherlands set up exercise wheels at two sites where wild mice live, and collected video for more than three years. They found that the median running speed of wild mice was lower than that of their captive counterparts, but wild mice reached higher maximum speeds. The animals were initially attracted to the equipment by food rewards, but continued to use the wheels when food was absent.

The study suggests that wheel-running is a voluntary behaviour, the authors say.

Proc. R. Soc. B 281, **20140210 (2014)**

MICROBIOLOGY

Fight and flight fortifies pathogen

A pathogen found in the mouth becomes more virulent by interacting with an innocuous oral bacterium.

The harmless species *Streptococcus gordonii* produces hydrogen peroxide as an antimicrobial, but the bacterium also makes lactate that benefits the pathogen *Aggregatibacter actinomycetemcomitans*, which causes gum disease. Marvin Whiteley of the University of Texas at Austin and his colleagues discovered that the pathogen either fights peroxide with a detoxifying enzyme, or escapes the site where the antimicrobial is produced.

These 'fight-and-flight' responses help the pathogen to move far enough away to be protected from the peroxide, but also to remain close enough to continue feeding on lactate. This results in an infection that is more severe than one caused by the pathogen alone, say the researchers.

Proc. Natl Acad. Sci. USA **http://doi.org/sv4 (2014)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SOCIAL SELECTION

Popular articles on social media

Price of knowledge prompts reflection

Researchers on social media have been hashing out the big question of just how much scientific information is out there, with many discussing a recent article on accessing that knowledge. Using data gleaned from Google Scholar and Microsoft Academic Search, researchers at Pennsylvania State University in University Park estimated that more than 114 million English-language scholarly works — journal papers, dissertations, master's theses, books and technical reports — exist as resources online. But much of the bounty can be had only for a price. Pollution scientist Andrew Singer of the NERC Centre for Ecology and Hydrology in Wallingford, UK, noted on Twitter that "only" 24% of the works were open-access. "That's quite a low number," he says. "It's disappointing to see how resistant we are to change."

Khabsa, M. & Giles, C. L. ***PLoS ONE*** 9, **e93949 (2014)**

Based on data from altmetric.com. Altmetric is supported by Macmillan Science and Education, which owns Nature Publishing Group.

NATURE.COM
For more on popular papers:
go.nature.com/alvao1

SEVEN DAYS *The news in brief*

RESEARCH

NIH ethics row

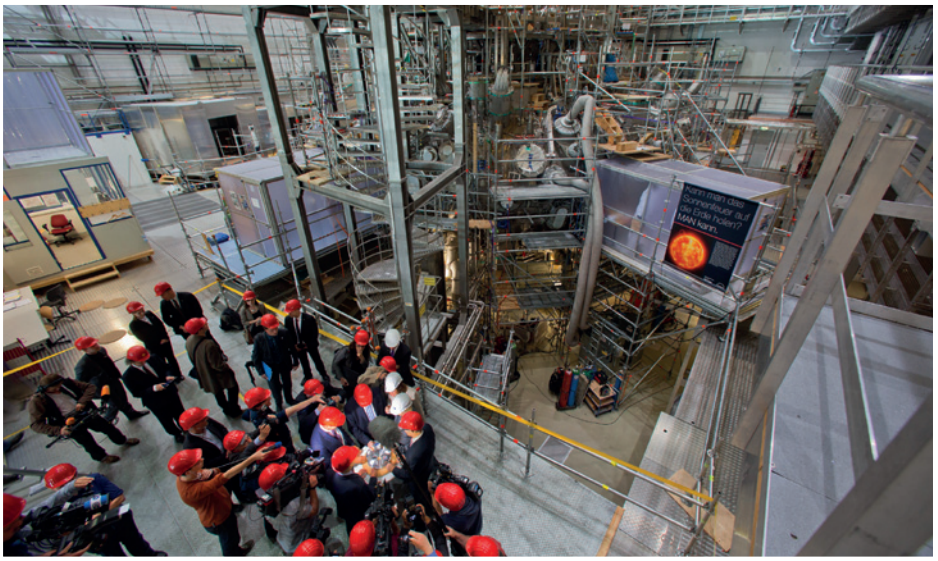
The US National Institutes of Health (NIH) may have interfered with an inquiry into one of its own clinical trials, according to internal e-mails released on 20 May by the watchdog group Public Citizen of Washington DC. Last year, the US Office for Human Research Protections (OHRP) criticized an NIH-funded trial of oxygen treatments in premature babies for failing to adequately inform parents of the risks. The NIH was allowed to edit drafts of the OHRP's report, the e-mails suggest. Public Citizen says that this compromises the investigator's integrity. See go.nature.com/3mmcuk for more.

Transgenic trials

Scientists in France are angry about the acquittal on 14 May of 54 activists who in August 2010 destroyed experimental transgenic grapevines in a field trial in the east of the country. In a joint statement released on 18 May, 12 research agencies and university organizations said that they had "serious concerns" about an appeal court's decision to throw out the case. The court said that the trials were illegal because the National Institute for Agricultural Research, which conducted them, had not proved that the vines would not cause health or environmental damage. See go.nature.com/do7mkp for more.

Study etiquette

Efforts to replicate psychology studies should follow a set of ground rules to foster good will, says Daniel Kahneman, a Nobel-prizewinning psychologist at Princeton University in New Jersey. In an open letter posted online on 20 May, he calls on would-be replicators to seek input on study design from the authors of the original report. Replicators should also provide authors with a full description of their study plans, Kahneman adds. His letter comes as the journal *Social Psychology* publishes a second tranche of replication efforts in the field, many of which failed to confirm the original findings.



JENS BUETTNER/EPA/CORBIS

Germany inaugurates 'stellarator'

The Max Planck Institute for Plasma Physics inaugurated its Wendelstein 7-X fusion experiment on 20 May, marking the end of installation and the beginning of preparations to start the machine in 2015. The €1-billion (US$1.4-billion) 'stellarator' project in Greifswald, Germany, features a wreath of twisted magnetic coils to confine a super-heated plasma of hydrogen isotopes. The much larger €15-billion ITER experiment being built in southern France will use a simpler doughnut-shaped magnetic-confinement device called a tokamak. Although notoriously difficult to design and build, a stellarator produces a more stable plasma than a tokamak.

Hurricane season

The Atlantic hurricane season this year is expected to be near normal, the US National Oceanic and Atmospheric Administration said on 22 May. A developing El Niño weather pattern in the equatorial Pacific Ocean is likely to suppress hurricane activity in the Atlantic, according to the agency. It predicts that 8–13 tropical storms will form during the season, which begins on 1 June. Of those, 3–6 are expected to develop into hurricanes.

Space-probe reboot

NASA has granted a group of citizen scientists permission to co-opt a defunct 1970s space probe. The unprecedented agreement, announced on 21 May, would permit the group's company, Skycorp of Los Gatos, California, to try to contact and take over control of the International Sun–Earth Explorer-3 (ISEE-3) when it approaches Earth in August. Skycorp wants to place the craft between Earth and the Sun, where it would reprise its role in monitoring space weather. The crowdfunded team plans to use a radio telescope at Morehead State University in Kentucky to command the probe. NASA launched the ISEE-3 in 1978 and shut it down in 1997.

FUNDING

German windfall

Eight months after Germany's federal elections, the government has finally outlined how it will make good on election promises to increase funding for research. On 26 May the government announced that research will receive €2.5 billion

(US$3.4 billion) out of an extra €6 billion to be injected into science and education in the 2014–17 legislative period. It is not yet clear how the windfall will be distributed between universities, research organizations and federally funded research programmes.

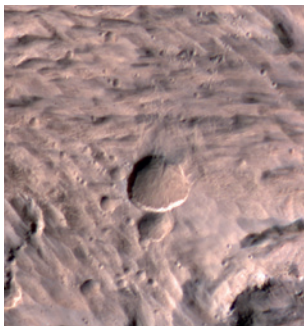
Vaccine plea

On 20 May, the GAVI Alliance — a global health partnership based in Geneva, Switzerland, that is seeking to improve access to vaccines in low-income countries — asked its donors, which include governments and charities, for an extra US$7.5 billion for 2016–20. For the period 2011–15, GAVI received $4.3 billion, $600 million more than requested. The group will ask donors to commit to pledges at a GAVI Alliance replenishment meeting next year in Germany.

EVENTS

Nuclear dump

In a bid to reduce the build-up of contaminated water beneath the destroyed Fukushima Daiichi nuclear reactors, Japan's Tokyo Electric Power Company (TEPCO) has begun to bypass upstream groundwater into the Pacific Ocean. On 21 May, it discharged the first tranche — 561 tonnes of stored groundwater.



NASA/JPL-CALTECH/UNIV. ARIZONA

Mars crater spotted

Some time between 27 March and 28 March 2012, this crater (**pictured**) formed on the surface of Mars — the biggest ever identified through 'before' and 'after' images. Scientists used archive images to confirm the formation date of the crater, seen here in an image taken by the Mars Reconnaissance Orbiter on 9 May.

POLICY

Review panel cut

The body that has overseen all gene-therapy research in the United States since 1974 is to have its activities drastically reduced. The US National Institutes of Health said on 22 May that the Recombinant DNA Advisory Committee, which it administers, will now publicly review only those gene-therapy trials that pose special risks and fail to receive adequate review by other governmental and institutional bodies. A US Institute of Medicine panel proposed this scale-back in December.

Covert crackdown

The US Central Intelligence Agency (CIA) will no longer use vaccination programmes to obtain information for its operations, according to a White House letter made public on 19 May. In 2011, the CIA was accused of mounting a fake hepatitis B vaccination campaign in Pakistan so that it could obtain DNA from children living in a compound later found to house Osama bin Laden (see *Nature* http://doi.org/bgbmgf; 2011).

US physics plan

High-energy physics in the United States must emphasize international collaborations to remain vibrant in the face of tight budgets, says a 22 May report from the US Department of Energy's Particle Physics Project Prioritization Panel. It stresses the need for the United States to remain a key player in the Large Hadron Collider at CERN, Europe's particle-physics lab near Geneva, Switzerland, and to stay involved in the proposed International Linear Collider in Japan. The panel also recommends recasting a proposed neutrino facility at the Fermi National Accelerator Laboratory in Batavia, Illinois, as an internationally funded effort. See go.nature.com/fo5b8f for more.

COMING UP

29–30 MAY
In Washington DC, scientists, policy-makers and environmental groups discuss how to restore ecosystems on large scales, with projects such as engineering wetlands.
go.nature.com/ra2zfd

4–15 JUNE
The United Nations holds climate talks in Bonn, Germany, to discuss progress under the Kyoto Protocol, and to continue negotiations towards a new treaty in Paris next year.
go.nature.com/iembtb

Chemical ban

Minnesota is to become the first US state to ban the sale of products containing triclosan, an antimicrobial agent found in soaps, body washes and other cleaning products. The ban was signed into law on 16 May. Triclosan is already under investigation by the US Food and Drug Administration (FDA) owing to safety concerns. The Minnesota law, which will come into effect on 1 January 2016, includes exceptions for products approved for consumer use by the FDA.

BUSINESS

Pharma deal off

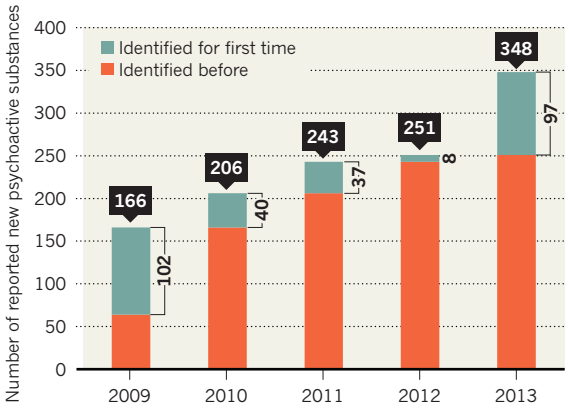
The US drug firm Pfizer has abandoned its pursuit of rival pharmaceutical giant AstraZeneca, it announced on 26 May. London-based AstraZeneca, together with some UK politicians and scientists, had fiercely opposed the proposal.

TREND WATCH

New illicit psychoactive substances are being developed at an "unprecedented pace" globally, according to a report released on 20 May by the United Nations Office on Drugs and Crime. Individual nations have banned some of these drugs, but none is under international control. More than two-thirds were synthetic cannabinoids, cathinones or phenethylamines. The latest numbers probably rose, in part, owing to a UN monitoring effort that was launched in June last year.

SURGE IN SYNTHETIC DRUGS

The United Nations drug agency reports that a cumulative total of 348 new psychoactive substances have been spotted since 2009.

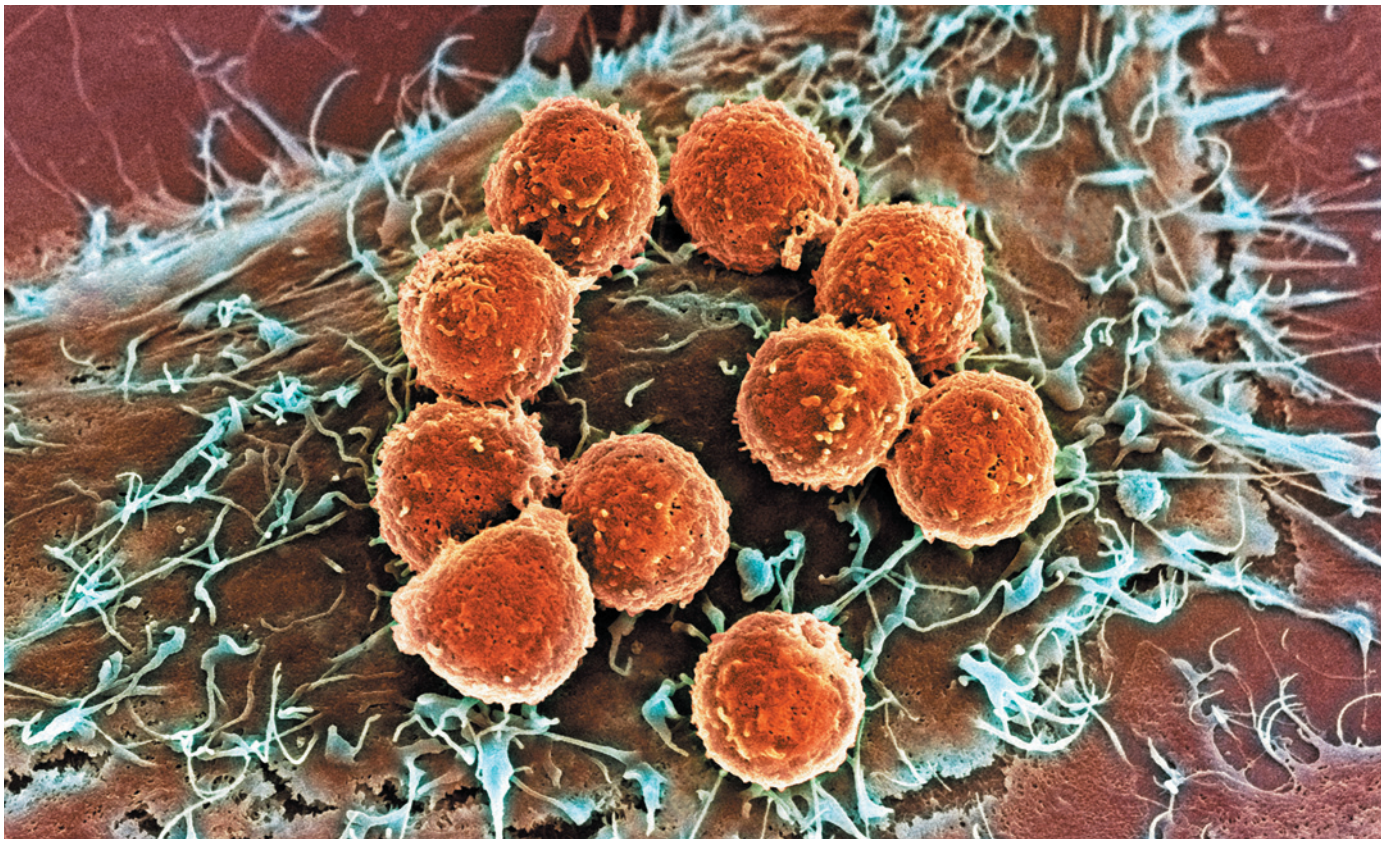


SOURCE: UNODC

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

STEVE GSCHMEISSNER/SPL



T cells (orange), which attach to and attack tumours, are fired up by immunotherapy treatments such as high-dose interleukin-2.

DRUG DEVELOPMENT

Old cancer drug gets fresh look

Refinements to disfavoured immune therapy offer safer, more effective treatment.

BY HEIDI LEDFORD

When Dave deBronkart was diagnosed with advanced kidney cancer in 2007, he learned about a treatment called high-dose interleukin-2 (IL-2) that fires up the body's immune system to fight the disease. The response rate was not great — tumours shrank in only about 15% of patients. And as many as 4% of people died from the treatment. But some of those who responded survived for years or even decades.

DeBronkart's prognosis was grim — only 8% of people with his disease survive for five years past diagnosis. He says that he was willing to risk what life he might have had left for the possibility of stopping his cancer's growth: "I said, 'Lethal side effects? OK.'"

Now, nearly seven years after his treatment, deBronkart's immune system continues to hold his cancer in check.

When researchers gather this week for the annual American Society of Clinical Oncology (ASCO) meeting in Chicago, Illinois, the spotlight will be on immunotherapies — a class of drugs that, like IL-2, kindle the immune system's ability to fend off cancer (see *Nature* **508,** 24–26; 2014). Attendees will hear the latest about a new generation of these drugs — particularly those targeting a protein called PD-1, which cancers use to fend off immune-system attack. Pharmaceutical companies are racing to bring these PD-1 inhibitors to market.

Other data to be presented at the meeting suggest that IL-2, the drug that saved deBronkart and the first cancer immunotherapy approved by the US Food and Drug Administration (FDA), may be on the verge of a revival after having fallen out of favour. Two years after he finished the therapy, deBronkart learned that because of IL-2's risks, three-quarters of eligible patients are never told that it is an option. "There are patients who are dying without ever getting a potentially curative treatment," says Steven Rosenberg, an oncologist at the ▶

US National Cancer Institute in Bethesda, Maryland. "It's a real problem."

IL-2 is a protein made by the body to spur the development of T cells in response to threats such as pathogens or cancer. Most cancer immunotherapies are designed to fire up T cells, but the effects can sometimes be deadly. Patients must be closely monitored while receiving high-dose IL-2 treatment, because it can cause skin rashes, short-term neurological disturbances and dangerous drops in blood pressure.

Given the risks and the difficulty of administering high-dose IL-2, patients and doctors often baulk at using the treatment, which is approved in the United States for use against advanced melanoma and kidney cancer. But Howard Kaufman, a tumour immunologist at Rutgers Cancer Institute of New Jersey in New Brunswick, argues that IL-2's bad reputation is outdated. Fatalities have declined as physicians have learned how to manage side effects, he says. And data from Kaufman and others suggest that more patients benefit from the drug than previously realized.

At the ASCO meeting this week, Kaufman will present an analysis of IL-2 treatment in people who received the therapy for kidney cancer between 2007 and 2012. Tumours shrank in only about 15% of patients, but they stopped growing in another 15%. Those people, said to have 'stable' cancer, lived longer than people whose cancers kept growing after treatment.

Even so, the therapy remains dangerous, warns Patrick Ott, a medical oncologist at Dana-Farber Cancer Institute in Boston, Massachusetts. "It can't be given at just any community hospital," he says. "You need a real expert team to administer this treatment."

"There are patients who are dying without ever getting a potentially curative treatment."

Ott and Kaufman are among researchers trying to make IL-2 therapy less toxic. At the meeting, they will describe their plans to test NHS-IL2, a fusion of IL-2 and an antibody that targets the DNA that is released by dying tumour cells. The antibody, developed by EMD Serono of Rockland, Massachusetts, directs IL-2 to tumour cells, helping to minimize its effects on healthy tissue. Another strategy, developed by a team led by John Frelinger, a cancer researcher at the University of Rochester Medical Center in New York, involves modifying IL-2 so that it is inactive until it encounters enzymes that are particularly abundant around tumour cells. "IL-2 was not meant to be expressed at high levels throughout the body," says Frelinger. "That's where the trouble with IL-2 therapy starts."

Others are experimenting with ways to expand the number of patients who respond to IL-2 treatment. In addition to activating cancer-fighting T cells, IL-2 stimulates regulatory T cells, which keep immune responses in check. Alkermes, a biotechnology company based in Dublin, is modifying the IL-2 protein so that it is less likely to activate regulatory T cells, and thus has the potential to generate a stronger immune response at lower doses. Because some of the toxic side effects of high-dose IL-2 seem to be independent of immune responses, it is possible that the approach will provide a more tolerable therapy.

Newer therapies, including PD-1 inhibitors, have milder side effects than IL-2. But even if those therapies are approved by the FDA, Ott thinks that IL-2 treatments of some stripe will still have a role, especially for people who do not respond to other treatments. "High-dose IL-2 has a track record of patients who have been disease free for 20 years," says Ott. "And we just don't know that yet with the new drugs." ■

OCEANOGRAPHY

US Arctic research ship ready to cast off

Long-awaited vessel Sikuliaq *joins an ageing fleet.*

BY ALEXANDRA WITZE

UAF/VAL IHDE

The latest addition to the US oceanographic research fleet will soon set off for its home port in Alaska.

A brand-new research vessel is buoying the hopes of US oceanographers. In the first week of June, the University of Alaska Fairbanks plans to take possession of the RV *Sikuliaq*, a US$200-million, 80-metre ship that is currently floating in the Great Lakes. It is the first research vessel built for the National Science Foundation (NSF) since 1981; polar scientists have been calling for a versatile ice-strengthened ship for four decades.

"People ask, 'Why should the Arctic have a special ship'? It's a special place," says Vera Alexander, a biological oceanographer at the University of Alaska who has been involved in the campaign to build an Arctic research vessel. Plans call for the *Sikuliaq* to spend much of its time examining the effects of shrinking sea ice and other climate-change impacts on northern ecosystems.

The *Sikuliaq*'s launch is particularly striking because most of the US oceanographic-research fleet faces a grim future. Many of the other 23 vessels are ageing, and there is little money available to replace them. Construction of the *Sikuliaq* was made possible only because, in 2009, then-NSF director Arden Bement chose to give the project $148 million from a government economic-stimulus package.

With its ability to navigate coastlines, ice-bound waters and the open sea, the *Sikuliaq* can explore a wide range of science questions, says Alexander. Biologists will be better able to track animal populations in places such

as the Bering Sea, one of the world's most biologically productive marine ecosystems. Geologists will be able to map the sea floor between Alaska and Siberia to reveal details about when the land bridge between the two was exposed, letting humans cross from Asia to North America. And chemical oceanographers will be able to track the spread of pollutants through once-pristine environments.

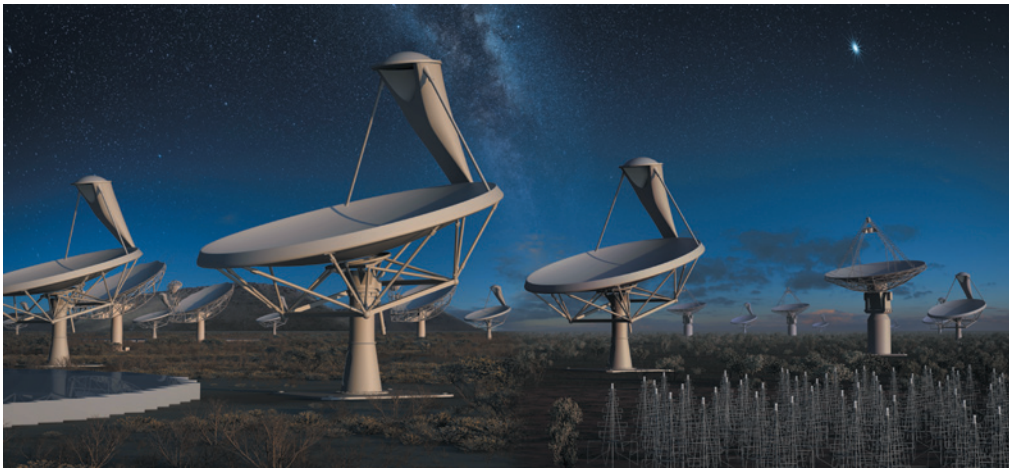
The *Sikuliaq* carries the latest oceanographic bells and whistles, including a high-tech boom that can be lowered over the side to deploy instruments such as sonar and oceanographic sensors. Unlike its predecessor, the now-retired *Alpha Helix*, the *Sikuliaq* has the ability to yank sediment cores up from the sea floor. A huge expanse of open deck space towards the bow will allow researchers to bring on board custom equipment, including autonomous underwater vehicles. "It's pretty amazing," says Michael Castellini, dean of the school of fisheries and ocean sciences at the University of Alaska.

The ship also has advanced navigation technology to improve manoeuvres through sea ice. That is important, because although its reinforced double hull allows it to plough through floating ice up to a metre thick and a year old, it cannot handle thicker, multi-year ice. (In line with the vessel's capabilities, *sikuliaq* is an Inupiaq word that means 'first-year sea ice that is safe enough to walk on'.)

Although construction delays have put it roughly a year behind schedule, "they didn't significantly alter when we wanted the first science to start", says Castellini. Plans call for the *Sikuliaq* to be at sea for some 270 days a year. One advantage to having a dedicated science vessel is that it will not be diverted for other purposes; research ships such as the US Coast Guard icebreaker *Healy* are occasionally called away to attend to emergencies such as delivering fuel to the icebound city of Nome, Alaska.

"People ask, 'Why should the Arctic have a special ship?' It's a special place."

Before it can do any science, the ship must pass final tests and be transferred from its builder, Marinette Marine of Wisconsin. After the University of Alaska team takes charge, the *Sikuliaq* will head out of the Great Lakes through the Saint Lawrence Seaway and then proceed southward along the US east coast. The ship will be based in Woods Hole, Massachusetts, for about a month while it undergoes shakedown tests at sea, and will then continue on, through the Panama Canal. It will do its first research in the deep waters of the tropical Pacific Ocean before heading north to reach its home port of Seward, Alaska, by February 2015. ■



SKA ORGANISATION

Data from the planned Square Kilometre Array of radio telescopes will require vast computing resources.

TECHNOLOGY

Cloud computing beckons scientists

Price and flexibility appeal as data sets grow.

BY NADIA DRAKE

Sometime in the next decade, the Square Kilometre Array (SKA) will open its compound eyes — roughly 2,000 radio dishes divided between sites in South Africa and Australia. The radio telescope will then begin staring into supermassive black holes, searching for the origin of cosmic magnetic fields and seeking clues about the young Universe.

Meanwhile, the telescope's engineers are struggling to plan for the imminent data deluge. The photons that will stream into the array's receivers are expected to produce up to 1 exabyte (10^{18} bytes) of data per day, roughly the amount handled by the entire Internet in 2000. Electricity costs for an on-site computing cluster big enough to process those data could total millions of dollars each year. So the engineers are investigating an increasingly common choice for researchers wrestling with big data: to outsource their computing to the cloud.

"No one's ever built anything this big before, and we really don't understand the ins and outs of operating it," explains SKA architect Tim Cornwell of the Jodrell Bank Observatory near Manchester, UK. He says that cloud systems — which provide on-demand, 'elastic' access to shared, remote computing resources — would provide an amount of flexibility for the project that buying dedicated hardware might not.

Such elasticity can also benefit projects that involve massively parallel data analyses, such as processing and aligning billions of DNA base pairs, or combing through hundreds of photos to identify specific zebras from their stripe patterns. It is also a boon to scientists who require bursts of computing power rather than sustained usage, as do researchers looking at seismic data in the aftermath of an earthquake.

"The rest of the year, when there are no earthquakes happening, they're just paying for storage," says David Lifka, director of the Cornell University Center for Advanced Computing in Ithaca, New York, which runs a computing-platform service called Red Cloud.

But the economics of cloud computing can be complex. An ongoing price war between major providers such as Google, Microsoft and Amazon Web Services has cut costs overall, but in many cases, sending data to the cloud, or retrieving them, remains much more expensive than storing them. Amazon's Elastic Cloud Compute S3 service charges US customers as much as US$0.12 per gigabyte for transfer from its servers, but no more than $0.03 per month to store the same amount of data.

This comes as a surprise to many researchers, according to a 2013 US National Science Foundation survey of 80 science projects that rely on the cloud (see 'Room to grow'). "Some cloud billing mechanisms are really opaque," says Daniel Perry, director of product and marketing for Janet, a private, UK-government-funded group near Oxford that is working to link British educational facilities to a shared data centre. "Unless you know what you're doing, you may find that you've run ▶

out of your grant in three months."

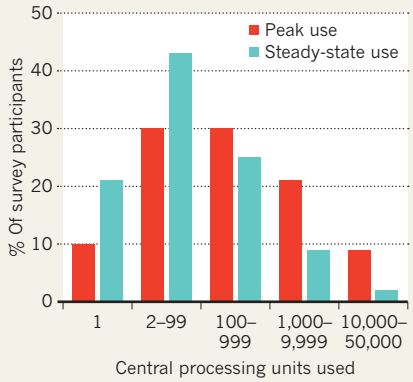
And costs aside, the cloud will probably never suit some computer projects, such as 'deep learning' networks that seek to mimic how the human brain learns. Adam Coates, a computer scientist at Stanford University in California who is involved in such work, says that these systems can require rapid information transfer between billions of connections — something not possible with the cloud. Instead, Coates relies on a dedicated on-site computing cluster. "Having that very high-speed communication is key," he says. "We want vast amounts of computation, but we don't really care about elasticity."

The cloud's dependability is also a concern, says Ken Birman, a computer scientist at Cornell. "It isn't known for being secure, and it isn't known for being extremely reliable." But not all researchers require foolproof data encryption or super-fast, reproducible computations.

For example, CERN, Europe's particle-physics laboratory near Geneva in Switzerland, assembled an in-house cloud to handle the data generated by the Large Hadron Collider. "The CERN data are public data, so we don't have any security concerns," says Tim Bell, who directs the centre's infrastructure and operating services. Instead, CERN focused on providing physicists with an efficient computing platform. "In the past, when they asked for physical hardware, they were waiting for weeks," Bell says. "Now they can ask for a virtual machine and get something in the time it takes to have a cup of coffee."

ROOM TO GROW

A 2013 survey of participants in 80 research projects powered by cloud computing reveals the flexibility it affords for handling data ebb and flow.



SOURCE: XSEDE CLOUD SURVEY

Universities are also getting into the cloud business. At Cornell, a subscription to Red Cloud costs $400 for 8,585 processing hours; for off-campus scientists, the same subscription is $640. Such on-campus services often appeal to researchers who are not ready for the do-it-yourself nature of commercial providers, which often requires expertise in programming, testing and debugging code. By contrast, Cornell cloud specialists are on site to help researchers using Red Cloud. "The thing you can't get with commercial clouds is hand-holding," Lifka says.

Meanwhile, companies such as Microsoft have set up cloud training specifically for academics, addressing issues such as data sharing and security, scientific reproducibility and how funding agencies may view the cloud. "A lot of the training and education content was tuned to a business audience. That meant the on-ramp for researchers was a bit more tricky," says Daron Green, senior director of Microsoft Research Connections. "We realized there was pretty much a latent demand within the research community." ■

MEDICAL RESEARCH

Biomedical institute opens its doors to physicists

The development is part of a growing trend to tap physics expertise.

BY ELIZABETH GIBNEY

From using soap bubbles when modelling cell division, to applying synchronized clocks to understand embryonic development, physics is becoming an increasingly effective tool for biologists. Now the field is to be a focus of a major new biomedical research hub in London, the Francis Crick Institute.

The £650-million (US$1.1-billion) centre — named after the co-discoverer of the structure of DNA, and a physicist-turned-biologist himself — will harness theoretical and experimental approaches from the physical sciences for medical research. When the institute opens in 2015, as much as one-fifth of its 1,250 staff will be physicists, chemists, mathematicians and engineers. They will be tasked with helping biomedical staff to understand why diseases develop and to find new ways to treat them.

The institute is a collaboration between the UK Medical Research Council (MRC), two charities — the Wellcome Trust and Cancer Research UK — and three London universities: Imperial College London, King's College London and University College London (UCL). The partners are already exploring the potential of bringing diverse fields together, with a workshop next week on astronomy and biomedical imaging. "When you see stars in the sky and you have to analyse their pattern and distribution and understand how they work, it's the same as looking down a microscope — the algorithms you use might well be very similar," says Jim Smith, a Crick board member and a director of the MRC's National

MORE ONLINE

TOP STORY

ANIMATED HEALTHCARE/SPL

Microbes use non-standard gene coding in nature go.nature.com/9a3zbc

MORE NEWS

- Healthy placenta plays host to bacterial flora go.nature.com/lmht8k
- How atoms form giant threesomes go.nature.com/yjeaxh
- NASA might end Spitzer telescope in bid to save other missions go.nature.com/geuakr

NATURE PODCAST

Protons' magnetism; antibiotic resistance; and fume-free stoves nature.com/nature/podcast

The interdisciplinary Francis Crick Institute will open in central London in 2015.

JUSTIN PIPERGER PHOTOGRAPHY/WADSWORTH3D

Institute for Medical Research, which will be subsumed into the Crick.

The term biophysics has in the past often been associated with the study of protein structure or ion-channel function. But the Crick is part of a movement that is taking physics in biology in a new direction. This could, for instance, involve modelling the emergence of shapes and patterns in biological systems across different scales, says Ewa Paluch, a cell biophysicist at UCL. "How do you go from molecules to cells, cells to tissues and organisms to groups of organisms? All these involve crossing scales, and are problems that can't really be understood without a model, and that's where physics comes in."

The approach could reveal how microscopic processes give rise to cell- and tissue-scale behaviours, and has been driven partly by improved imaging techniques and growth in the physics of 'soft matter', says Paluch, who will head a new Institute for the Physics of Living Systems at UCL, opening this autumn.

The application of physics in biology has been on the rise in the past decade, with increases in cross-discipline conferences, graduate courses and funding. Germany's Max Planck Society was among the first to support the approach when two of its centres, the Max Planck Institute of Molecular Cell Biology and Genetics and the Max Planck Institute for the Physics of Complex Systems, both in Dresden, began collaborating in 2002. The United States has also been a pioneer, says Alexander van Oudenaarden, a quantitative biologist at the Hubrecht Institute in Utrecht, the Netherlands. Until 2012, he was director of one of 12 Physical Sciences–Oncology Centers founded in 2009 by the National Cancer Institute, part of the US National Institutes of Health (NIH). Such was their success, for example in breast-cancer research, that last month the NIH opened a call for a new round of centres.

It might be argued that the emergence of the physics of biology is more of a renaissance than a new development, notes Paluch. Before the 'molecular revolution' of the 1950s, which refocused much biological research on genetics and molecular biology, physics played a bigger part in the field, she says — as demonstrated by Crick's involvement in the discovery of the DNA helix. But the shift sidelined physics.

Now the field can help biologists to handle the huge amounts of data they generate, image biological organisms in new ways and design materials, says Richard Treisman, director of Cancer Research UK's London Research Institute. "There are a huge number of issues where, like it or not, you're dragged kicking and screaming into the physical sciences," he says.

"It's not always easy to get a biophysics paper published. It's a problem we face constantly."

Such sentiments hint at a potential drawback in the coming together of disciplines — a clash of cultures. Biologists and physicists can sometimes speak different languages, so the communities at the Crick will need to learn to talk to each other.

The institute aims to smooth that process, says Smith. Whole groups from the physical-science fields will be able to take sabbaticals or secondments within the centre, and the facility has even been designed architecturally to encourage chance encounters that could stimulate ideas.

Biomedical funders have also recognized the benefits of engaging with physics. As part of its new research strategy, Cancer Research UK last month announced a £5-million fund dedicated to collaboration with the physical sciences, a decision it took after reviewing the effectiveness of such approaches.

This kind of initiative is welcome, says Paluch. Despite the rise in joint grants, convincing peer-review panels to fund interdisciplinary work is not straightforward, she adds. Nor is getting such work published, because many biology journals expect different kinds of mechanistic explanations from those usually put forward in physics. "It's not always easy to get a biophysics paper published. It's a problem we face constantly," she says. ■

JAGDEEP RAJPUT/ARDEA.COM

Modern chickens are descended primarily from the red junglefowl.

EVOLUTION

Chicken project gets off the ground

Effort aims to unravel the history of bird's domestication.

BY EWEN CALLAWAY

The meat and eggs of domestic chickens are a source of protein for billions. Yet how and when the birds were domesticated remains a mystery. The answers to these questions could reveal a wealth of information about the genetics of domestication, as well as human behaviour, and how we can improve our husbandry of the birds.

In a bid to learn more about the chicken and its lineage, the UK government is funding a £1.94-million (US$3.3-million) effort to determine how the chicken went from being a wild fowl roaming the jungles of southeast Asia several thousand years ago to one of the world's most abundant domesticated animals. The Cultural and Scientific Perceptions of Human–Chicken Interactions project — 'Chicken Coop' for short — will examine human history from the perspective of the fowl.

"No one ever considers chickens, which is a massive mistake," says Holly Miller, an archaeologist at the University of Nottingham, UK. She was one of two dozen researchers, from anthropologists to geneticists, who attended the second meeting of the five-month-old project at the University of Roehampton, UK, last week.

Another was Greger Larson, an evolutionary geneticist at Durham University, UK, who is a senior scientist on the project. He says that researchers studying domestication tend to overlook chickens in favour of other domesticated animals, such as dogs, cows and pigs.

But no domestic animal has been moulded and remoulded by humans as extensively as chickens, says Larson. The animals have been bred for eating, egg-laying and fighting. And in the case of one particularly vocal breed, the creatures have even been strapped to the masts of Polynesian boats to act as foghorns. "Chickens are polymaths," he says.

Larson, who studies DNA from the remains of ancient chickens, discovered last year that modern chickens can be deceptive. Previous studies have compared the DNA of modern chicken breeds with that of species of guinea fowl that contributed to the gene pool of early chickens, such as the red junglefowl[1,2]. The work identified variants in two genes that are common in contemporary chickens but not in guinea fowl.

One variant — when present in two copies — gives domestic chickens their familiar yellow skin and legs when they consume a diet rich in carotenoids; this trait is almost universal in European chickens. The other is a variant of the thyroid-stimulating hormone receptor gene (*TSHR*) that may alter the seasonal mating patterns of chickens and allow them to lay eggs all year round. It is universal in modern breeds such as Rhode Island Reds and broiler chickens.

Because these mutations are so common in contemporary chickens, Larson's team and others assumed that humans influenced these traits through selective breeding early in the course of domestication. But DNA from chickens recovered at archaeological sites across Europe, spanning the period from around 280 BC to AD 1800, has turned that idea on its head. In an analysis published last month, Larson's team reported that none of 25 ancient chickens would have had yellow legs, and that just 8 out of 44 birds carried two copies of the *TSHR* variant[3]. So even 200 years ago, chickens may have been very different from those we know today.

With the help of other Chicken Coop members, Larson is also trying to get to grips with the wider evolutionary forces that shaped modern chickens. He hopes to determine why, for instance, chickens have not been wiped out by disease. This might have been expected because their very rapid selection[2] — much of which has taken place since 1900 — should have led to inbreeding and, by whittling down immune genes, a reduced ability to respond to infections.

Other members of Chicken Coop are tackling different aspects of the bird's past. Miller plans to analyse the diet of ancient chickens, using chemical isotopes in their bones and egg shells, to reveal information on the resources available to the humans who kept them. Another team, at the University of Leicester, UK, will compare chicken bones from archaeological sites with bones of modern breeds. Known pathologies in today's birds can then be used to determine how diseases and breeding changed through time.

John Hutchinson, an evolutionary biomechanist at the Royal Veterinary College in London, thinks that a better understanding of the bird's history will help people to address some of the problems facing chickens and the poultry industry, such as avian influenza and leg weakness among broiler chickens. Research on ancient breeds could help us to "refresh the genetics" of broilers, he suggests. Last month, Hutchinson ran a conference, Towards the Chicken of the Future, to tackle such issues. "Science has got us into this problem through intense selection," he says. "It can maybe help us out of it." ■

1. Eriksson, J. *et al. PLoS Genet.* **4,** e1000010 (2008).
2. Rubin, C.-J. *et al. Nature* **464,** 587–591 (2010).
3. Flink, L. G. *et al. Proc. Natl Acad. Sci. USA* http://dx.doi.org/10.1073/pnas.1308939110 (2014).

CORRECTION

The News story 'Text-mining offers clues to success' (*Nature* **509,** 410; 2014) wrongly implied that SRI International is based in Arlington, Virginia. Although it has an office there, it is based in Menlo Park, California. It also should have said that IARPA is funding, not partnering with, SciCast, a project run by George Mason University and the American Association for the Advancement of Science.

Smoke from cooking fires, such as this one in Mumbai, India, kills millions of people a year.

VIVEK PRAKESH/REUTERS/CORBIS

Deadly dinners

BY MEERA SUBRAMANIAN

Polluting biomass stoves, used by one-third of the global population, take a terrible toll. But efforts to clean them up are failing.

After returning from her nine-and-a-half-hour shift as a security guard, Savita Satish Dadas begins plucking fenugreek leaves from their stems for dinner. She and her two children, along with three of their cousins, gather in a shed-like structure next to their house in the Satara District of Maharashtra, India. As goats and cows settle in for the night a few metres away, Dadas and the children sit down on a packed dirt floor around the family hearth.

Whisps of smoke rise up from their *chulha*, the Indian name given to a traditional cooking-stove fuelled by wood and other organic matter often gathered from the countryside. Dadas's stove, like several of her neighbours', is sculpted out of clay. But many make a rudimentary three-stone fire — a triangle of elevated points to support a pot — that humans have used for millennia. Dadas feeds roughly chopped logs into the stove and her hands shape moistened flour into *bhakri* bread, the rhythmic movement illuminated by the flickering flames.

With this simple daily act, Dadas shares a connection with more than one-third of the world's population, the three billion people who depend on solid biomass fuels — such as wood, animal dung, agricultural waste and charcoal — or coal for their cooking needs. In India, a nation that is rapidly developing in many ways, 160 million households — some two-thirds of families — still rely on such fuel for their primary cooking energy source. Globally, the percentage of households that use biomass has slowly and steadily decreased over the past three decades[1]. But because the world's population has been rising so quickly, the number of people using solid fuels is not declining, says Kirk Smith, an environmental-health scientist at the University of California, Berkeley,

who has studied the health implications of such cooking stoves for 30 years. "This is not going away."

And the urgency to transition billions of people around the world to cleaner forms of cooking has never been greater, in light of recent research revealing that emissions from traditional cooking-stoves pose a bigger threat than previously thought. Results from a global health study released earlier this year project that household air pollution from such fires causes more than four million premature deaths annually — more than one-quarter of them in India alone[2]. Earth's climate is also at risk from the smoke, which contains dark particles that absorb sunlight, alter atmospheric patterns and hasten glacial melting.

Environmental organizations, development groups and others have strived to solve the cooking-stove conundrum for decades, but momentum is finally gathering, thanks to the formation of the Global Alliance for Clean Cookstoves. This far-reaching public–private partnership was launched in 2010 by then US secretary of state Hillary Clinton. The Global Alliance has set a lofty goal of convincing 100 million households to adopt clean cooking-stoves by 2020, with an aim of eliminating deaths from cooking-stoves by 2030.

This massive effort — the most ambitious so far — is uniting specialists in fields as diverse as epidemiology, climate science, global finance and gender equality. It is part of a growing global effort that connects multinational energy companies, non-governmental organizations (NGOs), university design laboratories, governments and young, socially minded inventors. Simultaneously, new funding is flowing in from corporate social-responsibility initiatives, microfinance loans that provide credit to poor people, and the sale of carbon credits.

But all the efforts devoted to solving the problem have yet to make much of a dent. During three months touring the Indian states of Maharashtra and Tamil Nadu in late 2013 and early 2014, I interviewed dozens of women in their homes and found that improved cooking-stoves often sit unused in corners, broken or simply abandoned. My observations tally with those of field studies, which show that adoption rates of the new technologies remain as low as they have been for decades. The ongoing struggle is enough to make many researchers question whether it is truly possible to improve biomass cooking-stoves, and whether it might be better to direct efforts towards expanding access to proven technologies — such as gas stoves and electric cookers — that are already standard in the developed world.

It is time to move beyond age-old methods that cause so much pollution, both inside homes and out, says epidemiologist Kalpana Balakrishnan, director of the World Health Organization (WHO) Collaborating Centre for Occupational and Environmental Health at Sri Ramachandra University in Chennai, India. "If you want clean air anywhere, you don't want to be burning biomass in this configuration: open biomass burning."

COUNTING THE COST

The newest health data paint a stark picture of the impact of cooking with biomass. In March, the WHO estimated that 4.3 million people die annually from household air pollution caused by cooking with biomass and coal[3]. It is the greatest health risk in the world after high blood pressure, tobacco and alcohol[4], with more people dying from the incremental, ongoing inhalation of smoke from fires they ignite in their own homes than from malaria, tuberculosis and HIV/AIDS combined.

The new data more than double the WHO's 2004 estimate of the mortality rate from household air pollution. "This is not an energy issue," says Smith. "This is a health issue."

The data show that household air pollution from such fires causes acute lower respiratory infections, chronic obstructive pulmonary disease, cardiovascular disease and lung cancer[4]. Women and children, in particular, are often exposed to excessive amounts of small particles less than 2.5 micrometres in diameter, known as $PM_{2.5}$, which are considered the most dangerous to human health. A study[1] published by Smith and his colleagues this year that contributed to the WHO report[3] shows that Indian women cooking in households reliant on solid fuel are exposed to a mean 24-hour $PM_{2.5}$ concentration of 337 micrograms per cubic metre, more than ten times the WHO indoor air quality guidelines (see 'A burning issue').

Even before a match is struck, the stoves put women and girls at risk, because they are usually tasked with collecting the heavy loads of firewood or other materials. They also must often travel to remote locations to find fuel, making them vulnerable to sexual attacks. I have seen the signs of fuel collection across the south Asian landscape. Neat piles of slender branches are stacked high outside homes in Karnataka.

"Thirty years of research has really not produced a cost-effective way of burning wood."

Walls in Bihar are plastered with discs of drying cow dung impressed with petite handprints. Limber children scramble up a tree, hacking away branches with a machete in Punjab. A lone woman drags a 6-metre-long trunk down a sandy path in Tamil Nadu. At least there is fuel; India is more abundant in biomass than many places in Africa, where the situation is even more dire.

IN THE NEIGHBOURHOOD

Lata Kisan Kare, who lives near Dadas, says that she does not worry much about the smoke that pours out of the *chulha* standing just outside her front door. The pollution does not bother her, she explains matter-of-factly: "It goes up and away."

In reality, the smoke from Kare's fire is adding to the pollution in her village and beyond. In India, which now rivals China in terms of air pollution levels, one-quarter of the fine particulate matter in the ambient outdoor air originates from household cooking-stoves. Even people in households that have transitioned to liquefied petroleum gas (LPG) and other cleaner sources of fuel still have elevated pulmonary risk if their neighbours continue to cook with solid fuels, says Balakrishnan.

And the impact of such fires reaches around the globe. Evidence suggests that black carbon — sunlight-absorbing particles from cooking fires and other sources — are helping to weaken the Asian monsoon, melt mountain glaciers and speed up warming in the Arctic[5]. In 2013, a major assessment found that the black carbon emitted by sources such as cooking-stoves, diesel engines and agricultural fires is the second leading cause of climate warming after carbon dioxide emissions. In Africa and Asia, residential burning of solid fuels, including biomass and coal, accounts for a staggering 60–80% of black-carbon emissions.

The Global Alliance is trying to tackle these human and atmospheric problems through a range of activities, including improved monitoring and evaluation of cooking-stove programmes, and increased coordination between the hundreds of public, private, independent, non-governmental and funding entities across the 43 nations that are now partners under the alliance's umbrella. In 2012, the latest year for which data are available, partners distributed 8.2 million clean cooking-stoves.

But distribution is just one step in the move away from smoky fires. Households such as Kare's show how challenging it will be for the alliance to reach its goals, especially if the focus is primarily on improved biomass stoves. As part of a local corporate social-responsibility initiative, Kare received a free improved cooking-stove a few years ago from Cummins, a multinational corporation that operates an industrial megasite in the area. Five hundred stoves, each costing less than US$15, were installed in village homes by an NGO. Many, like Kare's, sit unused. The stove is a low-tech clay one designed for a cleaner burn, with a combustion chamber made of heat-resistant concrete and an air-intake hole for improved draft. Like dozens of stoves I saw in the Satara District, this one had a brick wedged in the air hole for fear that snakes or scorpions might mistake it for a lair.

Those who chose to use their stoves found that the design was flawed.

The wire holding the combustion chamber bricks together burned out quickly, causing the stoves to crumple under the weight of a pot, and many found that the fire still smoked too much.

Engineers have developed more sophisticated and sturdier designs, many of them portable, to prevent the kinds of problems seen in Kare's village. There are high-tech gasifier stoves, such as one from Philips, that rely on a rechargeable battery pack to run a fan for cleaner combustion. The Oorja also has a fan and burns pelletized field waste. Another option, BioLite, uses a thermoelectric generator to power a fan (as well as a USB charging port that might encourage mobile-phone-wielding husbands to buy improved stoves for their wives). And sleek units such as Envirofit and Prakti stoves use a natural draft to try to achieve a smoke-free fire.

But all of these seem to have limitations in the eyes of the users, who often reject them. Women told me that the stoves are too small to support a pot of bath water, or not hot enough to cook a *roti*, or flatbread. And many complained that they have to sit by the improved stoves and feed them continuously. With a conventional stove, they can just throw in a big log. Even those who do take to their new stoves face problems when the devices break; at present there is a very limited supply system in place for spare parts or repairs. The Global Alliance, along with companies such as Oorja, Envirofit and Prakti, are scrambling to put the necessary infrastructure in place, but the road remains uphill, and the best-performing stoves range from $50 to $80, far above the means of many of those who need them most.

Similar problems have plagued stove-improvement efforts for decades. A government cooking-stove programme in India reports distributing more than 30 million improved stoves between 1983 and 2002, but the World Bank and numerous researchers have criticized the programme, like many other stove initiatives over the years, for poor stove design, high programme costs, low adoption rates and lack of stove maintenance. Giving devices away has not seemed to work, and several stove designers told me that heavily subsidized programmes undermine the growth of a local market for stoves and spare parts that might help to buttress long-term use.

Although improved stoves produce less pollution, many people prefer traditional designs.

MEERA SUBRAMANIAN

TRADITION WINS

In late 2013, Smith prepared to revisit some of the villages he had first studied in the early 1980s. When we met up in Delhi just before his trip, he said, "I'm afraid that I'm going to see that nothing has changed." He was wrong. In the villages, people were chatting away on mobile phones, and many houses had electricity, satellite dishes and running water. But one thing had not changed: nearly all households still used *chulhas* for at least some of their cooking. "Development," Smith says, "has become unconnected with cooking."

Up in Smoke, a 2012 study by researchers at the Massachusetts Institute of Technology in Cambridge, highlighted some of the ongoing challenges[6]. A randomized controlled study in Odisha, India, identified no long-term improvements in health, fuel consumption or — the authors inferred — greenhouse-gas emissions in households that had been given a clean cooking-stove, primarily because the stoves were not being used. Although the devices had been distributed by the award-winning NGO Gram Vikas, they quickly fell into disuse, or were not maintained at a level that kept emissions low.

Smith criticizes the Up in Smoke study, saying that the introduced stove in question was known to be a poor one. His charge highlights the fact that no one knows how to define a 'clean' cooking-stove because there are no agreed standards for particulate emission from stoves.

Gautam Yadama, a professor of social work at Washington University in St. Louis, Missouri, and author of the book *Fires, Fuel & the Fate of 3 Billion* (Oxford Univ. Press, 2013), agrees that clean is a nebulous term. "What are the metrics?" he asks. "Who is calling them improved, and are they improved?" Efforts are afoot to address this issue. An International Organization for Standardization technical committee met in February in Nairobi to initiate development of standardized ways to test cooking-stoves, and the Indian government has also been busy developing labs that can approve certain cooking-stoves on the basis of thermal efficiency, as well as production of carbon monoxide and total particulate matter.

But none of the designs can get around some fundamental problems of burning biomass. Such fuels vary tremendously in terms of their moisture content and chemical composition, which makes it difficult to design an inexpensive stove that can burn cleanly in many situations. Moreover, users will invariably operate their stoves differently from a lab technician. And, no matter what the stove, biomass cannot pack as much energy as fossil fuels. "Thirty years of research has really not produced a cost-effective way of burning wood," Balakrishnan says. "Wood is not a calorific-enough fuel to burn very cleanly."

In some places, development is helping to make the question of adoption moot and hinting at a future without open fires. Many households are now using multiple types of cooking device, a strategy called 'stove stacking' that combines both modern and more traditional methods. This is especially evident in the southern state of Tamil Nadu, one of India's most developed areas.

When I step into Emily Teresa's house, above the Ladies and Gents Tailor Shop in the Krishnagiri District on a Saturday morning, a pressure cooker is whistling on an LPG stove in the kitchen, and a kerosene stove is stored under the counter. Teresa prefers the LPG, but limits her use of it — and the kerosene stove — according to how long her subsidized fuel supply lasts. To heat bathwater, she uses a traditional *chulha* outside. And, in another room, she keeps an Envirofit biomass stove that she purchased through a woman's cooperative that she belongs to. The NGO Integrated Village Development Project helped to bring 25,000 of those stoves to her district.

Her sister-in-law down the road also has multiple stoves, including an induction burner, an increasingly popular streamlined electrical unit that uses electromagnetic induction to transfer heat to pots. Where electricity is dependable, induction offers a stove that is much cleaner and more efficient, at a cost that is comparable to mid-range improved biomass stoves.

This stove-stacking by those at the bottom of the energy ladder is reminiscent of the way that those higher up segue seamlessly between gas ranges, microwaves and electric kettles, while a hot-water heater quietly does its work unnoticed. But amid the stack of options, the Envirofit stove is often the last one that Teresa and her sister-in-law reach for.

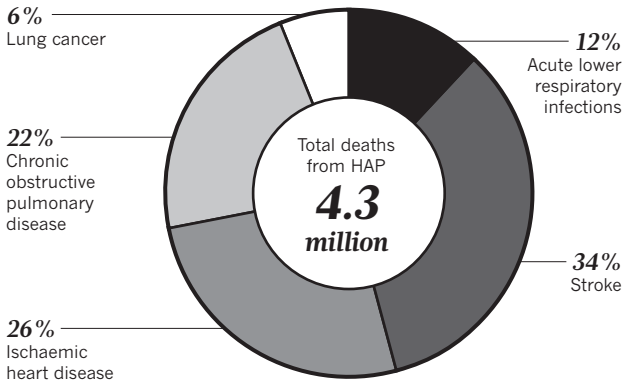
After decades of battling to get people to use improved

SOURCE: REF. 3

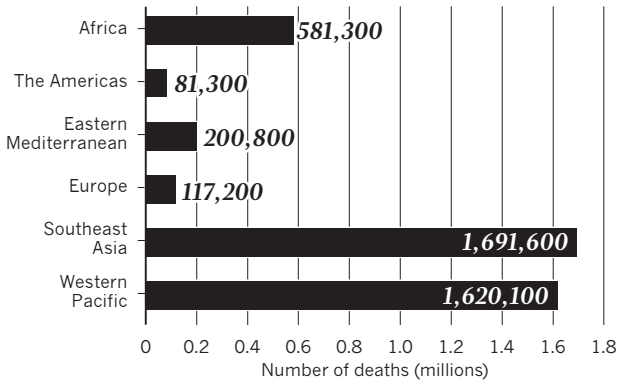
A burning issue

Nearly 3 billion people burn wood, dung and other types of biomass in open stoves to cook their food and heat their homes. The World Health Organization has estimated the number of deaths caused by household air pollution (HAP) from burning biomass and coal.

Cause of death from household air pollution in 2012



Total deaths attributable to HAP in 2012, by region



cooking-stoves, many researchers worry that such devices will never win over consumers and thus never achieve the desired health and climate gains. "My bottom line is that nothing works," Smith says. "The only thing we know that's ever worked is gas and electric."

Balakrishnan makes a moral argument against improved cooking-stoves, which still produce harmful amounts of pollutants compared with LPG or electric ones, powered by remote energy plants that commonly use fossil fuels. "Are you justified in saying that it's OK to be just a little bit better?" she asks. "If it's OK for 40% of the population to use fossil fuels, then why is not OK for the other 60% of the population? How can we have dual standards?"

ENERGY TRANSITION

Smith, Balakrishnan and others think that the answer may be for people to jump several rungs on the energy ladder, by-passing improved biomass stoves. It would be better, they suggest, for designers and policymakers to direct their efforts to helping more people transition directly to gas or electric stoves.

One of Kare's neighbours has done just that. She saved up enough to buy an induction stove and an LPG stove, and spends as much on refilling her subsidized gas cylinders, which she says last three months, as Kare spends buying just three-week's worth of wood fuel. Kare, too, would like the cleaner stoves, but the up-front costs are too high for her.

The rapidly changing energy landscape may be opening up new opportunities. Although Indians are apprehensive about the future of LPG subsidies, which are highly variable, many people are gaining access to new sources of alternative and renewable energy. In India and other developing countries, entrepreneurs are setting up decentralized electrical distribution systems fuelled by solar power, hydropower or biogas derived from agricultural waste (see *Nature* **507,** 154–156; 2014), which the world's rural poor have in abundance. Electric microgrids coupled with induction cookers could provide a means for millions to move away from polluting biomass stoves.

Even the Global Alliance of Clean Cookstoves acknowledges the advantages of abandoning biomass stoves of any type. "If people can afford to and are able to access the cleaner cooking technologies, including electric and LPG stoves, then that's wonderful from our perspective," says Sumi Mehta, director of programmes for the alliance. "But we also know that in the short term not everybody's going to be able to leapfrog to that." Of the three billion people burning biomass, at least one-third have little hope of moving up the energy ladder any time soon. For them, she says, the alliance will continue to invest in creating a cleaner biomass stove, no matter how challenging the job.

Putting the finishing touches to dinner as the children sit patiently beside her, Dadas has little time to worry about such issues. She dips a spoon into a Vicks container full of salt and adds the seasoning to the fenugreek greens that will accompany lentils and one *bhakri* bread for each of the six family members she is feeding tonight. Tomorrow is Christmas, which means a precious day off from the factory. As a Hindu, Dadas does not celebrate the holiday. When I ask her what she plans to do, she laughs sadly and says that she will use her axe and the extra time to go out collecting firewood. ■ SEE EDITORIAL P.533

Meera Subramanian *is a freelance writer in Cape Cod, Massachusetts. She travelled to India on a Fulbright-Nehru fellowship, which provided partial support for this piece.*

1. Smith, K. R. *et al. Annu. Rev. Public Health* **35,** 185–206 (2014).
2. Balakrishnan, K., Cohen, A. & Smith, K. R. *Environ. Health Perspect.* **122,** A6–A7 (2014).
3. World Health Organization *Burden of Disease from Household Air Pollution for 2012* (WHO, 2014) available at http://go.nature.com/smuctx.
4. Lim, S. S. *et al. Lancet* **380,** 2224–2260 (2012).
5. Bond, T. C. *et al. J. Geophys. Res.* **118,** 5380–5552 (2013).
6. Hanna, R., Duflo, E. & Greenstone, M. *Up in Smoke: The Influence of Household Behavior on the Long-Run Impact of Improved Cooking Stoves* (MIT, 2012).

COMPLEXITY ON THE HORIZON

SPL

A concept developed for computer science could have a key role in fundamental physics — and point the way to a new understanding of space and time.

BY AMANDA GEFTER

When physicist Leonard Susskind gives talks these days, he often wears a black T-shirt proclaiming "I ♥ Complexity". In place of the heart is a Mandelbrot set, a fractal pattern widely recognized as a symbol for complexity at its most beautiful.

That pretty much sums up his message. The 74-year-old Susskind, a theorist at Stanford University in California, has long been a leader in efforts to unify quantum mechanics with the general theory of relativity — Albert Einstein's framework for gravity. The quest for the elusive unified theory has led him to advocate counter-intuitive ideas, such as superstring theory or the concept that our three-dimensional Universe is actually a two-dimensional hologram. But now he is part of a small group of researchers arguing for a new and equally odd idea: that the key to this mysterious theory of everything is to be found in the branch of computer science known as computational complexity.

This is not a subfield to which physicists have tended to look for fundamental insight. Computational complexity is grounded in practical matters, such as how many logical steps are required to execute an algorithm. But if the approach works, says Susskind, it could resolve one of the most baffling theoretical conundrums to hit his field in recent years: the black-hole firewall paradox, which seems to imply that either quantum mechanics or general relativity must be wrong. And more than that, he says, computational complexity could give theorists a whole new way to unify the two branches of their science — using ideas based fundamentally on information.

BEHIND A FIREWALL

It all began 40 years ago, when physicist Stephen Hawking at the University of Cambridge, UK, realized that quantum effects would cause a black hole to radiate photons and other particles until it completely evaporates away.

As other researchers were quick to point out, this revelation brings a troubling contradiction. According to the rules of quantum mechanics, the outgoing stream of radiation has to retain information about everything that ever fell into the black hole, even as the matter falling in carries exactly the same information through the black hole's event horizon, the boundary inside which the black hole's gravity gets so strong that not even light can escape. Yet this two-way flow could violate a key law of quantum mechanics known as the no-cloning theorem, which dictates that making a perfect copy of quantum information is impossible.

Happily, as Susskind and his colleagues observed[1] in 1995, nature seemed to sidestep any such violation by making it impossible to see both copies at once: an observer who remains outside the horizon cannot communicate with one who has fallen in. But in 2012, four physicists at the University of California,

Santa Barbara — Ahmed Almheiri, Donald Marolf, Joseph Polchinski and James Sully, known collectively as AMPS — spotted a dangerous exception to this rule[2]. They found a scenario in which an observer could decode the information in the radiation, jump into the black hole and then compare that information with its forbidden duplicate on the way down.

AMPS concluded that nature prevents this abomination by creating a blazing firewall just inside the horizon that will incinerate any observer — or indeed, any particle — trying to pass through. In effect, space would abruptly end at the horizon, even though Einstein's gravitational theory says that space must be perfectly continuous there. If AMPS's theory is true, says Raphael Bousso, a theoretical physicist at the University of California, Berkeley, "this is a terrible blow to general relativity".

DOES NOT COMPUTE

Fundamental physics has been in an uproar ever since, as practitioners have struggled to find a resolution to this paradox. The first person to bring computational complexity into the debate was Stanford's Patrick Hayden, a physicist who also happens to be a computer scientist. If the firewall argument hinges on an observer's ability to decode the outgoing radiation, he wondered, just how hard is that to do?

Impossibly hard, he discovered. A computational-complexity analysis showed that the number of steps required to decode the outgoing information would rise exponentially with the number of radiation particles that carry it. No conceivable computer could finish the calculations until long after the black hole had radiated all of its energy and vanished, along with the forbidden information clones. So the firewall has no reason to exist: the decoding scenario that demands it cannot happen, and the paradox disappears.

Hayden was sceptical of his own result at first. But then he and Daniel Harlow, a physicist at Princeton University in New Jersey, found much the same answer for many types of black hole[3]. "It did seem to be a robust principle," says Hayden: "a conspiracy of nature preventing you from performing this decoding before the black hole had disappeared on you."

The Harlow–Hayden argument made a big impression on Scott Aaronson, who works on computational complexity and the limits of quantum computation at the Massachusetts Institute of Technology in Cambridge. "I regard what they did as one of the more remarkable syntheses of physics and computer science that I've seen in my career," he says.

It also resonated strongly among theoretical physicists. But not everyone is convinced. Even if the calculation is correct, says Polchinski, "it is hard to see how one would build a fundamental theory on this framework". Nevertheless, some physicists are trying to do just that. There is a widespread belief in the field that the laws of nature must somehow be based on information. And the idea that the laws might actually be upheld by computational complexity — which is defined entirely in terms of information — offers a fresh perspective.

➲ NATURE.COM
Read more about black-hole firewalls:
go.nature.com/husgnf

It certainly inspired Susskind to dig deeper into the role of complexity. For mathematical clarity, he chose to make his calculations in a theoretical realm known as anti-de Sitter space (AdS). This describes a cosmos that is like our own Universe in the sense that everything in it, including black holes, is governed by gravity. Unlike our Universe, however, it has a boundary — a domain where there is no gravity, just elementary particles and fields governed by quantum physics. Despite this difference, studying physics in AdS has led to many insights, because every object and physical process inside the space can be mathematically mapped to an equivalent object or process on its boundary. A black hole in AdS, for example, is equivalent to a hot gas of ordinary quantum particles on the boundary. Better still, calculations that are complicated in one domain often turn out to be simple in the other. And after the calculations are complete, the insights gained in AdS can generally be translated back into our own Universe.

> "The black hole's interior is protected by an armour of computational complexity."

INCREASING COMPLEXITY

Susskind decided to look at a black hole sitting at the centre of an AdS universe, and to use the boundary description to explore what happens inside a black hole's event horizon. Others had attempted this and failed, and Susskind could see why after he viewed the problem through the lens of computational complexity. Translating from the boundary of the AdS universe to the interior of a black hole requires an enormous number of computational steps, and that number increases exponentially as one moves closer to the event horizon[4]. As Aaronson puts it, "the black hole's interior is protected by an armour of computational complexity".

Furthermore, Susskind noticed, the computational complexity tends to grow with time. This is not the increase of disorder, or entropy, that is familiar from everyday physics. Rather, it is a pure quantum effect arising from the way that interactions between the boundary particles cause an explosive growth in the complexity of their collective quantum state.

If nothing else, Susskind argued, this growth means that complexity behaves much like a gravitational field. Imagine an object floating somewhere outside the black hole. Because this is AdS, he said, the object can be described by some configuration of particles and fields on the boundary. And because the complexity of that boundary description tends to increase over time, the effect is to make the object move towards regions of higher complexity in the interior of the space. But that, said Susskind, is just another way of saying that the object will be pulled down towards the black hole. He captured that idea in a slogan[4]: "Things fall because there is a tendency toward complexity."

Another implication of increasing complexity turns out to be closely related to an argument[5] that Susskind made last year in collaboration with Juan Maldacena, a physicist at the Institute for Advanced Study in Princeton, New Jersey, and the first researcher to recognize the unique features of AdS. According to general relativity, Susskind and Maldacena noted, two black holes can be many light years apart yet still have their interiors connected by a space-time tunnel known as a wormhole. But according to quantum theory, these widely separated black holes can also be connected by having their states 'entangled', meaning that information about their quantum states is shared between them in a way that is independent of distance.

After exploring the many similarities between these connections, Susskind and Maldacena concluded that they were two aspects of the same thing — that the black hole's degree of entanglement, a purely quantum phenomenon, will determine the wormhole's width, a matter of pure geometry.

With his latest work, Susskind says, it turns out that the growth of complexity on the boundary of AdS shows up as an increase in the wormhole's length. So putting it all together, it seems that entanglement is somehow related to space, and that computational complexity is somehow related to time.

Susskind is the first to admit that such ideas by themselves are only provocative suggestions; they do not make up a fully fledged theory. But he and his allies are confident that the ideas transcend the firewall paradox.

"I don't know where all of this will lead," says Susskind. "But I believe these complexity–geometry connections are the tip of an iceberg." ■

Amanda Gefter *is a freelance writer based in Cambridge, Massachusetts.*

1. Lowe, D. A., Polchinski, J., Susskind, L., Thorlacius, L. & Uglum, J. *Phys. Rev. D* **52,** 6997 (1995).
2. Almheiri, A., Marolf, D., Polchinski, J. & Sully, J. *J. High Energy Phys.* **2013,** 62 (2013).
3. Harlow, D. & Hayden, P. *J. High Energy Phys.* **2013,** 85 (2013).
4. Susskind, L. Preprint available at http://arxiv.org/abs/1402.5674 (2014).
5. Maldacena, J. & Susskind, L. *Fortschr. Phys.* **61,** 781–811 (2013).

COMMENT

ISSOUF SANOGO/AFP/GETTY

Unregulated sales of medicines in developing countries contribute to the rise in antimicrobial resistance.

An intergovernmental panel on antimicrobial resistance

Drug-resistant microbes are spreading. A coordinated, global effort is needed to keep drugs working and develop alternatives, say **Mark Woolhouse** and **Jeremy Farrar**.

Last month, the World Health Organization (WHO) produced a global map[1] of antimicrobial resistance, warning that a 'post-antibiotic' world could soon become a reality. In some ways, it already has.

Drugs that were once lifesavers are now worthless. Chloramphenicol, once a physician's first choice against typhoid, is no longer effective in many parts of the world. Strains of extensively drug-resistant tuberculosis (TB), methicillin-resistant *Staphylococcus aureus* (MRSA), multidrug-resistant *Escherichia coli* and *Klebsiella pneumoniae* are serious threats to public health. *Plasmodium falciparum* (the parasite that causes the most dangerous form of malaria) is developing resistance to all known classes of antimalarial drug, threatening the remarkable progress that has been made against the disease. HIV is increasingly resistant to first-line antiviral drugs. Every class of antibiotic is increasingly compromised by resistance, as are many antivirals, antiparasitic and antifungal drugs.

It could get worse: routine medical care, surgery, cancer treatment, organ transplants and industrialized agriculture would be impossible in their present form without antimicrobials. And the treatment of many infectious human and livestock diseases now relies on just one or two drugs.

Resistance has spread around the world. MRSA has spread between continents[2], as have resistant strains of TB, malaria, HIV and pneumococci. Genes conferring resistance to β-lactams — antibiotics used against a broad range of infections, including *E. coli* and *K. pneumoniae* — have spread to bacterial populations worldwide, probably originating in the Indian subcontinent[3]. Numerous drug-resistant malaria strains have spread from southeast Asia to Africa.

Antimicrobial resistance is a global problem that requires global solutions[1,4]. So far, the international response has been feeble. The WHO accepted only last month ▶

that antimicrobial resistance might fall within the remit of the International Health Regulations[1], which were implemented in 2007 to deal with events such as influenza pandemics. The regulations' extension to antimicrobial resistance would oblige the 196 signatory countries to carry out effective surveillance and timely reporting for outbreaks of resistance.

Better surveillance is essential. But it will not provide solutions; many calls to action on antimicrobial resistance have been made over the past 20 years, but there has been too little progress. The WHO missed the opportunity to provide leadership on what is urgently needed to really make a difference.

What is required is committed and coordinated action on the root causes of resistance: the misuse of antimicrobials, the paucity of development of new drugs and the lack of alternatives. Guidelines must be implemented to improve the use of existing drugs; the scientific and business worlds need incentives and a better regulatory environment to develop new drugs and approaches, and those working in both the animal and human sectors need education and incentives to help them to change their ways.

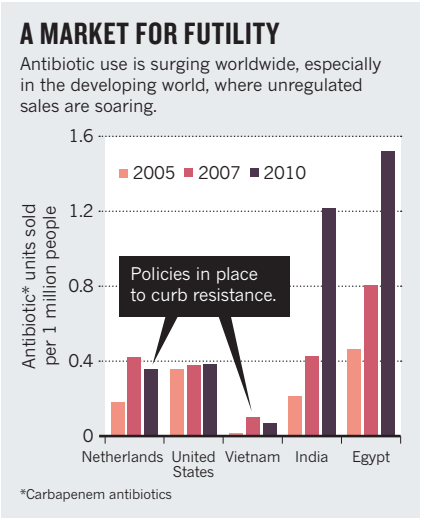
We call for the creation of an organization similar to the Intergovernmental Panel on Climate Change (IPCC) to marshal evidence and catalyse policy across governments and stakeholders.

USE AND MISUSE

Although all kinds of microbes evolve resistance, resistant bacteria are currently the greatest cause for concern. It is no coincidence that the nations with the strictest policies on antibiotic prescription (Scandinavian countries and the Netherlands) have the lowest rates of resistance. But in most of the developed world, clinical use of antibiotics has not declined, despite frequent calls to curtail overuse. In developing countries with rising incomes, consumption is surging; sales of even relatively expensive antibiotics increased fivefold in India and tripled in Egypt in 2005–10 (see 'A market for futility'). This growth is fuelled by unregulated, over-the-counter sales of antimicrobials of all kinds.

In the United States, antibiotic usage in humans is matched by that in farm animals, mainly as growth promoters. The European Union banned the use of antibiotics as growth promoters in animals in 2006, but the situation is little better. As industrialized agriculture expands, notably in Asia, animal antibiotic usage will continue to grow.

Mitigating resistance will require coordination across sectors. Physicians, pharmacists, veterinarians, patients and farmers all contribute to the overuse of antimicrobials. All have a part to play in using them more intelligently. However, changing practices in the hospital, clinic or farm is not easy. The onus is on countries that are major producers and consumers of antimicrobials — especially the United States and European nations, and increasingly India and China — to introduce policies that promote best practice.



Currently, national efforts are patchy and disconnected. The United Kingdom last year published a five-year strategic plan to combat resistance (see go.nature.com/ideq6t), although with no new money attached. Vietnam aims to combat resistance through its VINARES project[5], but most countries have no such programmes. The United States is still debating how to reduce the use of growth promoters in animals. Regional initiatives such as the European Antimicrobial Resistance Surveillance Network are yet to be replicated elsewhere. Controls that do exist are often weakly implemented or are no more than voluntary guidelines.

RESISTANCE IS NATURAL

Most of the antibiotics in use today, from penicillin to carbapenems, originated in soil. Long before they were used as medicines, soil microbes were producing antibiotics, and bacteria were evolving resistance to these natural compounds. This has been happening for perhaps billions of years[6] on a massive scale: there are at least 50 tonnes of bacteria for every person on the planet[7].

Humans became involved with the manufacture of antibiotics on an industrial scale only in the 1940s. Today, 20 tonnes of antibiotics are produced every hour, contributing to a global industry that is worth more than US$30 billion a year. We are now in a race against evolution; new antimicrobials are deployed and, often within a few years, resistance develops. Factory-produced antibiotics are presenting bacteria with a type of chemical attack that they have overcome many times before.

Between 1983 and 1992, 30 new antibiotics were approved by the US Food and Drug Administration. From 2003 to 2012, the number was just seven. Why? Because there are too few incentives and too many regulatory barriers for the commercial sector to invest what is needed for the development of new antimicrobials[8]. Drug development is risky, and antibiotics do not generate as much revenue as drugs for chronic conditions do. Drug companies find that research in other diseases is a better return on investment.

SOURCE: REF. 4

A GLOBAL APPROACH

In many ways, antimicrobial resistance is similar to climate change. Both are processes operating on a global scale for which humans are largely responsible. In antimicrobial resistance, as in climate change, the practices of one country affect many others.

One key difference is that, for climate change, technologies exist to produce energy without burning fossil fuels, and investments and incentives will make them practical and affordable. Alternatives to antimicrobials — such as probiotics, prebiotics or phage therapy — are still, at best, experimental[4]. More research on alternatives is urgently needed, coupled with efforts by industry, academia and governments to market them in a scalable way.

There have also already been internationally agreed, evidence-based targets for cutting carbon dioxide emissions. There are no global targets for reducing antimicrobial use and no real understanding of how to set them. We do not even know what, if any, level of antimicrobial usage will be sustainable in the long term.

The threat of anthropogenic climate change led to the creation in 1988 of the IPCC. Despite its limitations, the panel is arguably the most successful attempt in history to empower scientific consensus to inform global policy and practice.

Another useful precedent is the Montreal Protocol on Substances that Deplete the Ozone Layer, the first universally ratified treaty in the history of the United Nations. Faced with clear data that the ozone layer, which protects Earth from ultraviolet radiation, was under threat, governments agreed on a timetable to phase out ozone-depleting chemicals. The protocol, which came into force in 1989, is considered the most successful global environmental treaty, and has led to the shrinking of the ozone hole.

We believe that similar global approaches should be attempted to address problems in public health. There is a need for a powerful panel to marshal the data to inform and encourage implementation of policies that will forestall the loss of effective drugs to resistance, and to promote and facilitate the development of alternatives — a panel akin to the IPCC, and the analogous Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services founded in 2012. An intergovernmental panel on antimicrobial resistance (IPAMR) must have

the same firm foundation on the best available science and potentially an even stronger mandate for action.

From the outset, the IPAMR needs to avoid simply restating the problem. It must move rapidly to an agenda that includes identifying key knowledge gaps and how to fill them; assessing viable short- and long-term solutions; evaluating barriers to implementation; and setting out road maps for sustainable control of disease-causing microbes. It could, for example, support studies to investigate dosing regimes that stall resistance, coordinate incentives for developing new types of antimicrobial and set targets for prescriptions and animal use.

To have any chance of achieving these objectives, the IPAMR must be trusted and free of vested interests. It will need to involve a broad range of experts, encompassing clinical and veterinary medicine, epidemiology, microbiology, pharmacology, health economics, international law and social science. It will need technical, financial, industrial and political support from governments and agencies including the WHO, the World Organisation for Animal Health, the World Trade Organization and the United Nations, as well as from representatives of producers and consumers of antimicrobial drugs. Above all, it will need strong, independent leadership.

Creating an effective IPAMR will be a huge undertaking, but the successful global campaign to eradicate smallpox, led by the WHO, demonstrates that a coordinated, international response to a public-health threat can work. The attempt must be made — otherwise, the massive health gains made possible by antimicrobial drugs will be lost. ■

Mark Woolhouse *is professor of infectious disease epidemiology in the Centre for Immunity, Infection & Evolution at the University of Edinburgh, UK.* **Jeremy Farrar** *is director of the Wellcome Trust, London, UK.*
e-mails: mark.woolhouse@ed.ac.uk; j.farrar@wellcome.ac.uk

1. World Health Organization *Antimicrobial Resistance: Global Report on Surveillance 2014* (WHO, 2014).
2. Harris, S. R. *et al. Science* **327,** 469–474 (2010).
3. Vernet, G. *et al. Emerg. Inf. Dis.* **20,** 434–440 (2014).
4. Laxminarayan, R. *et al. Lancet Inf. Dis.* **13,** 1057–1098 (2013).
5. Wertheim, H. F. L. *et al. PLoS Med.* **10,** e1001429 (2013).
6. D'Costa, V. M. *et al. Nature* **477,** 457–461 (2011).
7. Whitman, W. B. *et al. Proc. Natl Acad. Sci. USA* **95,** 6578–6583 (1998).
8. Cooper, M. A. & Shlaes, D. *Nature* **472,** 32 (2011).

ILLUSTRATIONS BY PETE ELLIS/DRAWGOOD.COM

Bring microbial sequencing to hospitals

Analysing bacterial and viral DNA can help doctors to pick effective drugs quickly, says **Sharon Peacock**.

A patient goes to her doctor with fever, cough and night sweats. Rapid tests confirm the diagnosis of tuberculosis and hint at multidrug resistance. But to suggest the optimum drug combination, as many as eight weeks of laboratory testing are required — a timescale dictated by the slow growth rate of the pathogen (*Mycobacterium tuberculosis*). In the meantime, the doctor must make an educated guess about which medicines to prescribe, increasing the risk of ineffective treatment and spread of infection.

Yet it would take less than a week to

sequence a culture of *M. tuberculosis* and to detect mutations that indicate which drugs the bacteria are resistant to. My colleagues and I demonstrated[1] this in a research setting last year using a sputum culture from a tuberculosis patient. We have also shown that whole-genome sequencing can detect resistance of other pathogens to carbapenem antibiotics, drugs reserved to treat the most serious infections. Although a range of genetic mechanisms can confer resistance, sequencing is informative in all species tested so far, including *Acinetobacter baumannii* and *Klebsiella pneumoniae*, which can infect the most vulnerable hospital patients[2]. Sequence information can also be used to confirm outbreaks and help to bring them to a close.

Although technology for microbial sequencing has existed for years, it is yet to help patients on a routine basis. Now that pathogen genomes can be completed in less than a day, the time is right to begin using them to control and treat serious infections, at least in the developed world. This will take two developments: the introduction of sequencing into local diagnostic laboratories and the creation of automated tools to interpret newly sequenced genomes. Both are more a matter of will than of innovation.

USEFUL DATA

A baby born in 1930 in the United States had a life expectancy of around 60 years; someone born today has a life expectancy of about 80. By some estimates, antibiotics have contributed as much as a decade to that jump in lifespans. Now the rise of antimicrobial resistance threatens these gains[3]. In a post-antibiotic era, even minor infections could prove fatal. Worse, without effective antibiotics, most medical practices, including routine surgery, emergency operations, transplants and chemotherapy, will be less safe.

Microbial sequencing could help physicians to know which antibiotics will be effective for their patients. Although it cannot reveal previously undocumented forms of resistance, a genome sequence can simultaneously reveal mutations and acquired genes that bestow resistance to many antibiotics. By contrast, real-time polymerase chain reaction (PCR), a more limited form of DNA analysis, can detect no more than a handful of known resistance markers in a sample. Our 2013 sequencing analysis[1] showed that a patient with tuberculosis was infected with a mixture of two strains, both resistant to more than a dozen antibiotics including some that are not routinely evaluated (see 'Resistance on the rise').

"In a post-antibiotic era, even minor infections could prove fatal."

Several research groups have used sequencing to trace outbreaks of multidrug resistant pathogens in hospitals. For example, a retrospective genomic investigation[4] of a 2011 outbreak of carbapenem-resistant *K. pneumoniae* at the US National Institutes of Health Clinical Center in Bethesda, Maryland, helped researchers to reconstruct the most likely transmission routes and revealed a network that included asymptomatic patients and contaminated medical equipment. This could guide targeted interventions including cleaning and sterilizing procedures.

NEED FOR SPEED

Human genomes capture the limelight in most discussions about the benefits of and barriers to bringing sequencing into health care. Yet pathogen sequencing represents a quick win. Bench-top sequencers that cost around US$125,000 can complete several bacterial genomes in a day, at a cost of around $150 per sample — about twice as much as running some commercial PCR tests that detect resistance to one drug at a time.

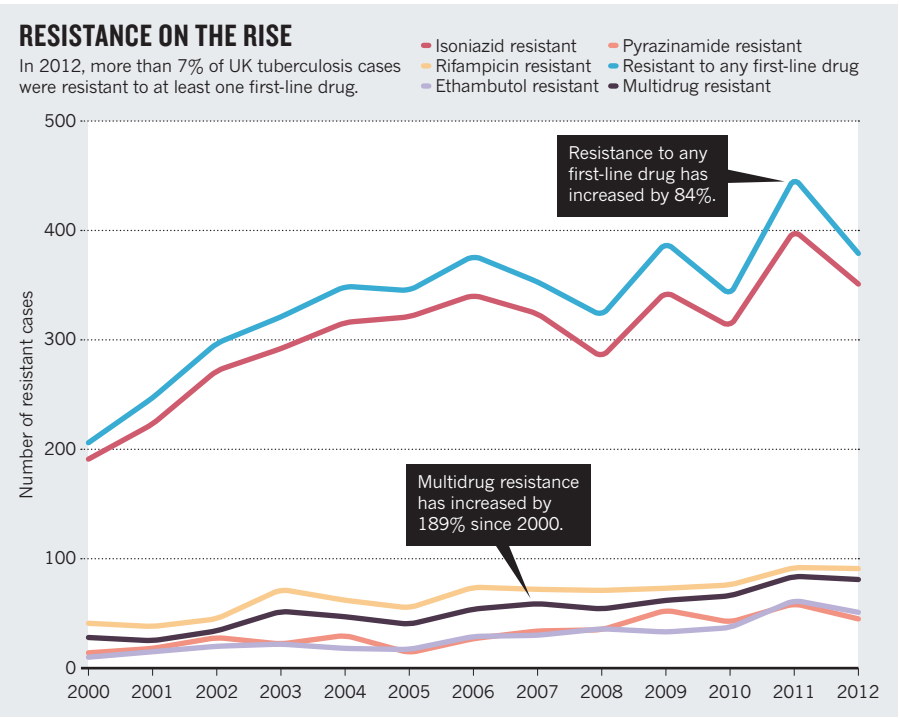
People working to bring human sequencing to the clinic generally aim to establish large, centralized facilities to control costs, maintain quality and facilitate data sharing. A 15-day turnaround or better is the target by the end of the 5-year Genomics England programme, which aims to sequence 100,000 human genomes by 2017, according to Mark Caulfield, the programme's chief scientist. This timescale is tolerable because the data rarely relate to conditions that are immediately life-threatening.

With infectious diseases, however, swift answers really matter. Suspected outbreaks can be confirmed and nipped in the bud. Within an hour of receiving results, clinicians could tailor treatments of hospitalized patients, providing the right drug at the right dose and route of administration. This precision will help more people than just the patient. Infections that are cured quickly are less likely to spread. It also means using fewer ineffective drugs, reducing the selective pressure for resistance, and leading to fewer pointlessly disrupted gut microbiomes.

Therefore, microbial sequencing should be done as close to the patient as possible, avoiding delays from packaging and shipping. This is feasible. Clinical microbiology laboratories are in place across the developed world, and have a track record of embracing new technologies. An example is the uptake of mass spectrometry starting five or so years ago to identify pathogens based on tell-tale microbial peptides. Diagnostic laboratories have also adopted real-time PCR, which entered clinical diagnostic laboratories about a decade ago.

Existing mechanisms to develop standard operating procedures and accreditation are applicable to microbial sequencing. Diagnostic laboratories already have tightly controlled systems to collect, track and process samples. They have defined turnaround times, they can link test results to patient and infection-control information, and they use an information system that ensures patient confidentiality.

Sequencing for HIV is already established



SOURCE: PUBLIC HEALTH ENGLAND

in diagnostic laboratories. It is performed intermittently over patient's lives to detect the emergence of viruses resistant to therapy. Sequencing for other viruses is also likely in the near future. With a raft of new drugs to treat hepatitis B and hepatitis C entering the market, detecting resistance will be required for hepatitis drug trials and clinical care.

AUTOMATED ANALYSIS

The major barrier to bringing microbial sequencing to the clinic is the lack of automated analysis tools. Once laboratory technicians have sequenced a pathogen, they need to convert the data into information that can be understood by non-specialists.

A probable scenario is that registered clinical users will access a web-based system. This is the model used for HIV interpretation. The HIVdb Program at Stanford University in California accepts user-submitted sequences of key genes and predicts levels of resistance to commonly used drugs. Analysis of other pathogens is likely to follow suit: submitted genetic data will be compared to a reference database of known variants and how they affect drug susceptibility. To be effective, reference databases will require considerable upkeep: new variants must be incorporated and tested for their effects on susceptibility.

The development of a single interpretation pipeline for all microbial sequence data will be impossible at first. For example, detecting resistance in bacteria that cause an acute infection is different from doing so in viruses that cause long-term infection. The type of analysis needed to predict resistance is also different from that required to investigate an outbreak. Specifically, instead of scanning for established resistance markers, newly sequenced genomes have to be compared against others to assess whether a related genome had been observed previously within a hospital, and if so when and where.

Academic and commercial competition is likely to lead to the emergence of a suite of excellent tools. However, data will be more powerful if they can be brought together. The collation of microbial sequences generated in a country could provide national surveillance on the emergence and spread of antimicrobial resistance. Because antimicrobial resistance is a global problem, further linkage will be needed to produce an international database, an unprecedented opportunity to detect resistance and new disease threats. It could also serve as an early-warning system for the emergence of strains not controlled by existing vaccines.

FINDING FUNDING

Although the potential benefits of an international network are clear, funding is less so. The translation of microbial sequencing into the clinic has largely been supported by short-term research funding to validate technology and timelines. Sustaining a long-term international programme will require a different funding model.

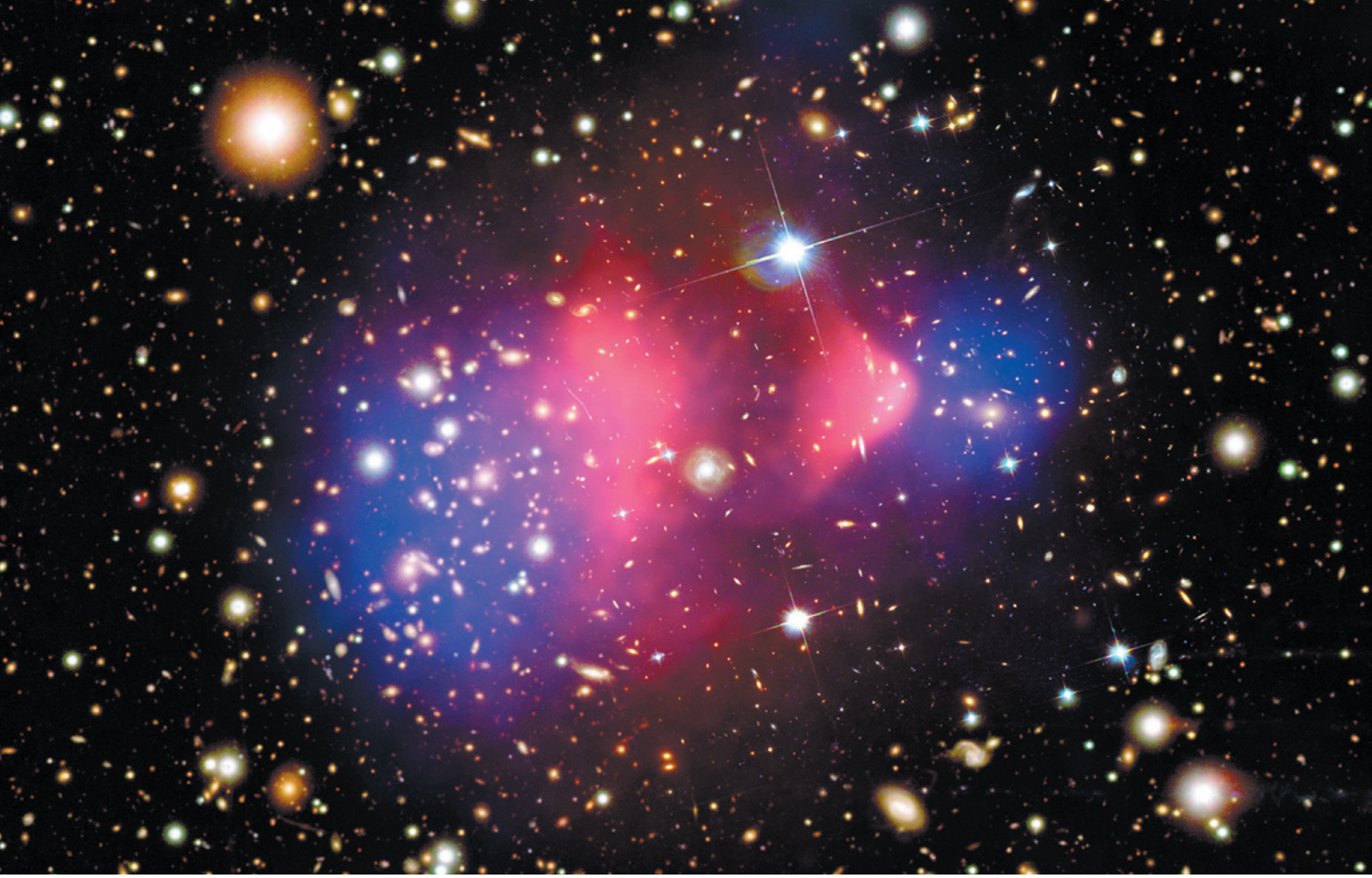
Launched in 2011, the Global Microbial Identifier is an initiative to create a genomic epidemiological database to identify micro-organisms, which could be used to detect outbreaks, antimicrobial resistance and emerging pathogens. Experts from more than 30 countries have signed on, and the effort is attracting interest from funders, governments and academia.

An essential issue will be deciding who is permitted to access microbial genome data generated during clinical care. Sharing data with the research community and pharmaceutical companies would be useful for developing drugs and evaluating interventions, but will require safeguards to protect the identity of individuals. Analysis of microbial genomes could be used to try and reveal who gave which disease to whom, which is especially sensitive for sexually transmitted diseases. The use of viral gene-sequence data in the law court to prove that HIV transmission occurred directly between two individuals is unsound and unwelcome[5]. Failsafe systems that prevent data from being accessed by unauthorized individuals are obligatory.

The application of sequencing technology to microbial genomes will improve patient care and enhance public health. The feasibility and economics are clear. To reap these benefits, the logistics must be established. With automated analysis tools and laboratory procedures in place, sequencing can help patients, stall outbreaks and help to stave off the post-antibiotic era. ■

Sharon Peacock *is professor of microbiology at the University of Cambridge, UK, and an honorary consultant microbiologist for Public Health England and the Cambridge University Hospitals NHS Foundation Trust.*
e-mail: sjp97@medschl.cam.ac.uk

1. Köser, C. U. *et al. N. Engl. J. Med.* **369,** 290–292 (2013).
2. Reuter. S. *et al. JAMA Intern. Med.* **173,** 1397–1404 (2013).
3. World Health Organization *The Evolving Threat Of Antimicrobial Resistance: Options For Action* (WHO, 2012).
4. Snitkin, E. S. *et al. Sci. Transl. Med.* **4,** 148ra116 (2012).
5. Bernard, E. J. *et al. HIV Med.* **8,** 382–387 (2007).



In the 1E 0657-56 galaxy cluster, ordinary matter is shown in pink. Blue indicates most of the mass of the cluster — presumed to be dark matter.

NASA

COSMOLOGY

Matter and mixology

Francis Halzen is exhilarated by an account of the hunt for the particles of dark matter.

The juggernaut that is precision cosmology has uncovered a strange Universe: 7 parts dark energy, 2.5 parts dark matter and 0.5 parts hydrogen and helium gas (with traces of other chemical elements). The stars and neutrinos, microwave photons and supermassive black holes that constitute the rest do not add up to very much. *The Cosmic Cocktail* by astrophysicist Katherine Freese tells us how we got to this bizarre recipe, what it might mean and what could come next.

The book itself is a cocktail, mixing science and autobiography. When the flood of facts and explanations threatens to overwhelm the reader, Freese delights us with vignettes of her personal life. So we get accounts of her favourite places to drink and dance in New York City or to ski in Aspen, Colorado, as well as entertaining stories about the work of dark-matter hunters she has met and collaborated with, such as Bernard Sadoulet and Juan Collar. Her portrait of astroparticle pioneer David Schramm, who in 1997 tragically died while flying his plane home for Christmas, is touching. It all works because, in prose as in life, Katherine Freese is never boring. Hers is an insider's view of how cosmology has been transformed since the 1960s, from a niche science to a discipline that pushes the intellectual frontier of physics — a development in which Freese has had an active and prominent role.

Ever the master mixologist, Freese gives us a sequence of clear and accessible introductions to the key concepts of cosmology and its observational techniques. The approach is never superficial and often quantitative. The treatment of nucleosynthesis is a gem, laying out how the nuclear physics of the early Big Bang left us with a Universe whose ordinary matter is made mostly of hydrogen and helium. Even the ratio of these gases is computed, at the price of introducing two of the very few equations in the book.

The Cosmic Cocktail: Three Parts Dark Matter
KATHERINE FREESE
Princeton University Press: 2014.

Much of the narrative focuses on the hunt for dark matter. The nineteenth-century French astronomer Urbain Le Verrier proposed the existence of a new planet, Vulcan, to explain strange wanderings in the orbit of Mercury. He was wrong, it transpired, because he was using an outdated theory — Newtonian gravity rather than Einstein's general theory of relativity. One might wonder whether the strange apparent composition of the cosmological cocktail hints, in turn, at the demise of Einstein's masterpiece. But Freese strongly argues that dark matter is not the Vulcan of general relativity, and that the search for its particle constituents will eventually find success.

The centrepiece of the book is the search for dark matter's particles. Starting with Fritz

Zwicky in the early 1930s, astronomers have built a conclusive case for the existence of dark matter. Its presence is inferred from its gravitational pull on stars in the Milky Way, and on whole galaxies in galaxy clusters. As with the neutrino, we are not aware of its existence: we do not feel or see it. In fact, neutrinos with an appropriate mass were once thought to make up dark matter, but this attractive idea — neutrinos do exist, after all — does not create the Universe that our telescopes observe.

"One might wonder whether the strange apparent composition of the cosmological cocktail hints at the demise of Einstein's masterpiece."

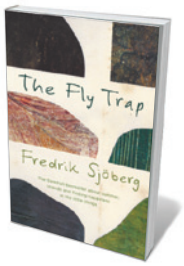
Also like neutrinos, dark-matter particles do not interact with normal matter, or not very much. More than two decades of intensive searching have not yet brought a discovery, but have instead measured an upper limit on the cross-section, or probability, of a dark-matter particle interacting with ordinary matter.

Freese dramatically describes the highly competitive race to the smallest cross-section limit, introducing the main players and their wide-ranging instrumentation. Because instruments are made from ordinary matter, they are not generally affected by the passage of a dark-matter particle. For even the slightest chance of success, one must build detectors of enormous size or develop increasingly imaginative detection techniques — preferably both. In most detectors today, the hope is that a dark-matter particle will bounce off a nucleus in a detection medium, producing recoil and generating a tiny amount of light or heat, which detectors can then harvest to infer the particle's existence. It is not a race to the bottom, however, and a discovery — with its guaranteed Nobel prize — could be around the corner at any time.

The Cosmic Cocktail is an excellent primer for the intrigued generalist, or for those who have spent too much time in particle-physics labs and want to catch up on what cosmologists are up to. The book will undoubtedly inspire students, too. It exposes the two great secrets about science that solar-neutrino guru John Bahcall often mentioned in his lectures: science is addictive, and unpredictable. It proceeds with detours, dead ends, false alarms, missed opportunities and surprises. And one day, someone might run into dark matter. ■

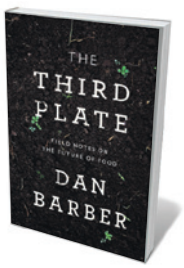
Francis Halzen *is the principal investigator of the IceCube South Pole neutrino observatory, and Hilldale and Gregory Breit professor of physics at the University of Wisconsin-Madison.*
e-mail: francis.halzen@icecube.wisc.edu

Books in brief

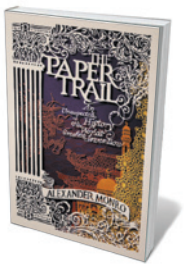


The Fly Trap
Fredrik Sjöberg (Translated by Thomas Teal) ALLEN LANE *(2014)*
A shimmering and elusive grace pervades Fredrik Sjöberg's evocation of his life and work as a hoverfly expert. This Swedish best-seller deftly interweaves three threads: the intricate business of fieldwork tracking and trapping flies in the Stockholm archipelago; the life of sawfly expert, art collector and sometime crackpot René Malaise; and a wealth of cultural allusions. Sjöberg is mesmerizing, whether describing the psychology of collecting, hoverflies as "superb impostors" or the moment a maple blooms in greenish glory — when "everything flies, absolutely everything".

The Sick Rose: Disease and the Art of Medical Illustration
Richard Barnett THAMES AND HUDSON *(2014)*
Can a panoply of horrors become a thing of beauty? Richard Barnett would have you think so. He has collected scores of appalling, if brilliantly rendered, illustrations from eighteenth- to twentieth-century medical textbooks and other sources, to explore a period in which medicine moved conclusively away from its unscientific past. In his view, images of syphilitic sores and horrifically eroded faces can and should be considered legitimate art as well as historical artefacts. A fascinating book, albeit only for those with strong stomachs.



The Third Plate: Field Notes on the Future of Food
Dan Barber PENGUIN *(2014)*
Mapping the provenance of meat and produce — the "farm to table" approach — is now part of the Western food-sustainability paradigm. Lauded chef Dan Barber calls for a more radical shift. His finely orchestrated agricultural model incorporates crop biodiversity and rotation, mixed land use, diverse livestock diets, ethical fishing and a realignment of US eating habits towards 'whole food' cuisine, including 'nose-to-tail' consumption of animals. An inspirational agronomic model — but given the dominance of industrial agriculture in the United States, it may smack of the utopian to some.



The Paper Trail: An Unexpected History of the World's Greatest Invention
Alexander Monro ALLEN LANE *(2014)*
Paper may be derided as a waste of trees, and as dead as the dodo in our digitized world. But, as Alexander Monro reminds us in this erudite history, it has been the base layer of world culture. Monro traces paper's trajectory from its second-century-AD origins in China to its passage through Eurasia, the Maghreb and on to world dominance. From Islamic scientific tracts to Copernicus's 1543 *De Revolutionibus*, paper, as Monro eloquently shows, has filled the supremely important role of placing "truth in the reader's hands".

The Next Crash: How Short-Term Profit Seeking Trumps Airline Safety
Amy L. Fraher CORNELL UNIVERSITY PRESS *(2014)*
The 700 million people flying in the United States this year might face unexpected turbulence, argues former naval aviator Amy Fraher. Her history of the US aviation industry — including hundreds of interviews with professionals — reveals cost-cutting and lax oversight in recent years. Short-term gain is trumping long-term safety: in 2000–12, carriers earned US$2 trillion, yet risk management is not keeping pace with rapid shifts in the industry. **Barbara Kiser**

FOX

In *Cosmos*, Neil deGrasse Tyson discusses life, the Universe and everything.

Q&A Neil deGrasse Tyson

The space crusader

US astronomer Neil deGrasse Tyson, director of New York's Hayden Planetarium, currently hosts the television series Cosmos — an *update of Carl Sagan's 1980 show — broadcast in 181 countries and 45 languages. As it winds down, Tyson talks about the rich mix of science and pop culture, the 'neurosynaptic snapshot' of public responses to his tweets, and his momentous meeting with Sagan.*

Were you surprised by the reaction to *Cosmos*?
Here's what I was shocked by, delightfully shocked: the publicity went way beyond the entertainment pages. There were articles in all the science sections, and news journalists were writing about it. It became more than a television phenomenon: it became a cultural phenomenon. It was all over the blogosphere.

Why was the time right to remake *Cosmos*?
The receptivity of our culture to science and to a major science programme has been steadily growing. One of the most popular sitcoms on the major US networks now is *The Big Bang Theory*, about scientists at the California Institute of Technology in Pasadena. You cannot be the number one sitcom unless many people other than just geeks watch you. Nowadays, a scientist on a television show is a fully fleshed-out character and has the same problems as everybody else — they want to fall in love. That's how you know there's some mainstreaming going on.

Why did you personally decide to do it?
I felt that I could do *Cosmos* in a way no one else could — that I could express the passion that anyone in astronomy has for the Universe in a way that relates to pop culture and storytelling.

What do you hope people will take away?
I would count the show a success if people looked at the world around them differently, and understood that it is knowable through the methods and tools of science — and that those methods and tools have shaped the civilization that we now take for granted. If they get that much, they will come to understand how science works and that it is not a satchel of facts but a process, a way of decoding the secrets of nature.

What was the hardest part of the preparation?
My biggest challenge was knowing that the show was scripted, but delivering it as though it wasn't. I've never had to do that in my life. As an educator, as an academic, I'm used to talking, but for *Cosmos* I had a speech coach. I would read the script and the coach would say, "If you say this word in this other way, then it will sound more like you." I had to deliver the lines the way an actor does. You're not thinking they're reading a script, but that they're feeling that emotion on the spot.

How has *Cosmos* changed you as a communicator of science?
I learned how collaborative such an ambitious project needed to be for it to rise to the heights that it has. And I learned what a gaffer and key grip do. Also, I became comfortable communicating with a camera lens rather than a room of people. I value the instant feedback from a person's facial expression, so this was harder than I had imagined.

Tell me about your meeting with Carl Sagan.
On my application to Cornell University in Ithaca, New York, I said that I was deeply into the Universe and astronomy. The admissions officer forwarded my application to Sagan, who worked at Cornell. He sent me a letter of invitation to come up and visit. He gave me a tour of his lab and a copy of his book, *The Cosmic Connection*. I still have that book, with his inscription: "For Neil, with all good wishes to a future astronomer". At the end of the visit, he drove me to the bus station and it began to snow a little harder. So he wrote his home phone number on a piece of paper and said if the bus can't get through, call me and spend the night with my family. I was just a 17-year-old kid and this is something I never forgot.

What's next for you?
There's a pet project that I'm returning to — my show *StarTalk Radio*, now in its fifth season. It has an inverted construct: it's about science but my guests are hardly ever scientists. Most are from pop culture; their fans follow them to this show and then hear them talk about science. It's an experiment, a new way of trying to reach the public that targets people who never knew how much they'd love learning about space and science. Among the people I've interviewed is the actor Morgan Freeman. You learn that he looked at the stars in the night sky on the porch in Mississippi when he was growing up and wondered where we all came from. He wanted to communicate something about that and now he's host of the US Science Channel documentary series *Through the Wormhole*. This is someone who's not a scientist but who's been touched by science.

How do you use social media?
I will send a tweet and watch what people say about it. It's an instantaneous neurosynaptic snapshot of people's reactions to even subtle points that are elucidated in the tweet. When I give a public talk, I feel like I already know the people there because there's communication through Facebook; social media is access to people's spontaneous emotions and thoughts. ■

INTERVIEW BY RON COWEN

Correspondence

Stop Madagascar's toad invasion now

Asian common toads (*Duttaphrynus melanostictus*) have begun to invade Madagascar, threatening the biodiversity of its unique fauna. Time is short, so we are issuing an urgent call to the conservation community and to governments to prevent an ecological disaster.

The first reported sighting of *D. melanostictus* on Madagascar was on 26 March in Toamasina. We collected six adult toads from a swamp in the humid eastern region, six kilometres from Madagascar's largest seaport. More were spotted nearby, suggesting that they arrived from Asia inside shipping containers, as they have elsewhere (see F. Kraus *Alien Reptiles and Amphibians*; Springer, 2009).

The region provides ideal resources and climate for the toad's spread into the island's interior, and *D. melanostictus* can range up to an elevation of 1,800 metres. Surveys are now being directed by the biodiversity organization Madagasikara Voakajy to identify the extent of invasion and develop a programme of eradication.

The species poses a significant risk to native fauna, given its life-history characteristics, the evolutionary naivety of the native fauna to toad toxins, and the damage caused in Australia and elsewhere by its relative, the cane toad (*Rhinella marina*).

Without swift eradication of *D. melanostictus*, the ecological consequences of an invasion include poisoning and decline of vulnerable native predators (birds, mammals, reptiles), the spread of amphibian diseases, and the secondary effects of food-web disruption. Potential impacts on humans include loss of domestic animals, contamination of drinking water and transmission of parasites in areas with poor sanitation.

Jonathan E. Kolby* *James Cook University, Townsville, Queensland, Australia.*
jonathan.kolby@my.jcu.edu.au
**On behalf of 11 co-signatories (see go.nature.com/4ataw3 for full list).*

US patient network safeguards data

Concerns about privacy have led the US National Patient-Centered Clinical Research Network (PCORnet; see go.nature.com/d9eaee and *Nature* **508,** 432; 2014) to adopt a slightly different approach from the UK National Health Service's care.data programme (see *Nature* **507,** 7; 2014), which has similar goals.

Launched by the Patient-Centered Outcomes Research Institute, PCORnet aims to improve patient outcomes by connecting patients, clinicians, researchers and health systems.

PCORnet will not use a centralized database to pool information from multiple networks. Rather, it will leave identifiable patient data behind the firewalls of the 29 participating networks, and distribute programming code that allows the same analysis to be run within each network. In most cases, only aggregated results are transferred for pooling and reporting, which avoids privacy or security risks.

Each PCORnet partner network must engage all its stakeholders in governance. The authority to make final policy decisions on privacy and security and to participate in multi-network studies is vested locally. All research uses of data will be approved and overseen by institutional review boards.

In some cases, individual informed consent will be deemed necessary. In low-risk research, such as observational studies that do not alter treatment choices, networks can favour other forms of communication with patients, especially if obtaining individual consent would render the study unfeasible. Such decisions will also be made locally, and with patient participation.

These measures aim to ensure that PCORnet's 'big data' approach will answer patients' questions about their care while safeguarding the privacy of their personal health information.

Joe V. Selby *Patient-Centered Outcomes Research Institute, Washington DC, USA.*
jvselby@pcori.org

Tackle pollution from solar panels

There is a downside to China having become the largest producer and consumer of solar energy (J. A. Mathews and H. Tan *Nature* **508,** 319; 2014). The rapidly expanding manufacture of solar photovoltaic products is risking serious environmental pollution.

According to Greenpeace and the Chinese Renewable Energy Industries Association, some two-thirds of the country's solar-manufacturing firms are failing to meet national standards for environmental protection and energy consumption. The production of polysilicon and silicon wafers for solar panels creates dangerous by-products, in particular silicon tetrachloride and hydrofluoric acid, which are being discharged into the environment after inadequate waste treatment (see go.nature.com/mhtayz; in Chinese).

For example, in 2011, fluoride concentrations in the Mujiaqiao River near a solar-panel factory in Haining City, eastern China, were more than ten times higher than permitted, killing fish and raising concerns about human health.

Another source of pollution is the careless disposal of used solar-panel equipment, which includes battery waste containing lead, cadmium, antimony and sulphuric acid (see H. Wang and J. Nima *Qinghai Soc. Sci.* **5,** 58–60; 2007).

Improved waste treatment, environmental monitoring and education are essential to avoid the undesirable impacts of these otherwise valuable technological advances.

Hong Yang *University of Oslo, Norway.*
Xianjin Huang *Nanjing University, China.*
Julian R. Thompson *University College London, UK.*
hongyanghy@gmail.com

Industries depend on biodiversity too

Andrew Moss and colleagues lament that raising the public's awareness of biodiversity will not necessarily change behaviour (*Nature* **508,** 186; 2014). A fresh perspective might: the public needs to recognize that biodiversity is crucial to a variety of core industries and not just the province of conservationists.

Biodiversity is not only about vertebrates and flowering plants, as is popularly believed: invertebrates and microbes account for at least 90% of all species. The genetic, metabolic, physiological and chemical diversity of these species underpins primary industries such as agriculture, grazing, forestry and fisheries.

Many biodiversity elements help to provide crops, timber, seafood and other necessities through ecosystem services. For example, microbes naturally regulate nitrogen and phosphorus in agricultural soils, and wild pollinators increase crop yields. It is therefore ironic that these primary industries so often constitute a threat to biodiversity conservation efforts.

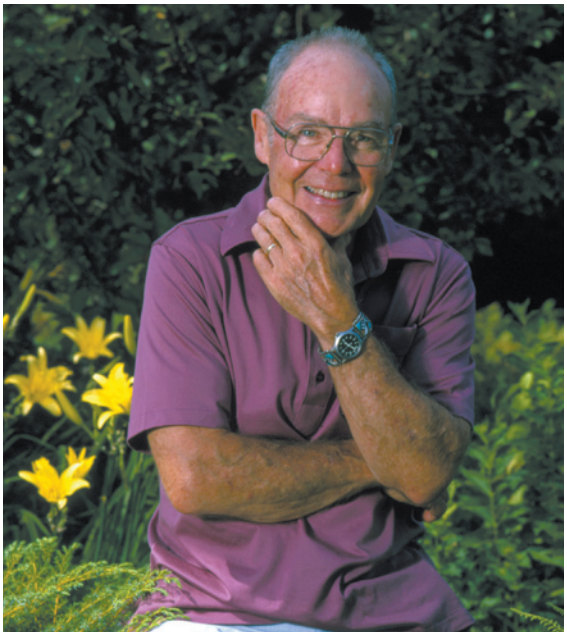
Promoting a popular vision of biodiversity that embraces all kinds of species could lead to conservation being taken more seriously by the public, and by those economic sectors that act as though biodiversity is not important to them.

Andrew J. Beattie *Macquarie University, Sydney, New South Wales, Australia.*
Paul R. Ehrlich *Stanford University, California, USA.*
andrew.beattie@mq.edu.au

Douglas Coleman

(1931–2014)

Biochemist who revealed biology behind obesity.



FRANÇOISE GERVAIS/JACKSON LABORATORY

More than 40 years ago, when few thought that obesity needed an explanation beyond a lapse in willpower, Douglas Coleman proposed that body weight and appetite are regulated by a then-undiscovered hormonal system.

Coleman was born in Stratford, Canada, in 1931, as the challenges of the Great Depression were reverberating around the world. When he was young, his parents lost their jobs in furniture manufacturing. The extended family lived together in one house, surviving by hunting for rabbits and squirrels and by farming in their backyard. As circumstances improved, his father opened a business repairing radios and refrigerators. Rather than joining him, Coleman went to study chemistry at McMaster University in Hamilton, Canada. There, he met a fellow chemistry student, Beverly Benallick. They married and shared a life together for the next 50 years.

After receiving a PhD in biochemistry from the University of Wisconsin in Madison, Coleman was recruited to the Jackson Laboratory in Bar Harbor, Maine. There, he studied two stocks of mutant mice. Both carried recessive mutations that produce marked obesity. The *ob* stock was severely obese and mildly diabetic, whereas the *db* stock was less obese and severely diabetic. Coleman and his colleagues used cross-breeding to study both mutations in the same inbred strain, and revealed that the strain differences vanished when either mutation was carried on an identical genetic background. In a series of important papers, Coleman showed that neither gene acts alone, but is modified by others.

When mutations in different genes produce the same effect, they often function in the same pathway, and Coleman began to probe how the genes work together. With developmental geneticist Elizabeth Russell and others, he launched in the 1960s a now-iconic set of experiments, in which he stitched together the skin and muscle of living mice of the different strains so that they shared a circulatory system. The *ob* mice that were surgically joined to normal or *db* mice ate less and lost weight, so Coleman concluded that *ob* mice lacked a circulating factor (provided by the conjoined partner) that suppressed food intake and body weight. Similarly, normal mice paired with *db* mice starved to death. Coleman suggested that *db* mice lacked a receptor to detect the weight-suppressing factor and so overexpressed it, producing levels so high that conjoined mice sensitive to the factor stopped eating.

Coleman spent years pursuing his hypothesis. Although he was not able to isolate the 'satiety factor' from circulating blood, his confidence in his conclusions never wavered; he explained that he had spent years trying to disprove his hypothesis and was never able to do so.

It took more than two decades for Coleman to be proved correct. Inspired by his experiments, my laboratory at Rockefeller University in New York City cloned the *ob* gene in 1994 and identified leptin, a hormone secreted by fat that regulates food intake by binding to its receptor. The leptin receptor is encoded by the *db* gene and expressed in the hypothalamus, a region of the brain known to control basic functions. Coleman also lived to see the US Food and Drug Administration approve the hormone as a treatment for severe diabetes associated with lipodystrophy, an often devastating disease resulting from an absence of fat tissue.

The discovery of leptin has led to the elucidation of a robust system that maintains relatively constant levels of fat stores. When fat mass falls, so too do leptin levels in the blood, stimulating appetite and suppressing energy expenditure. When fat mass increases, leptin levels rise, suppressing appetite. These findings have transformed obesity research, overhauling the scientific view of fat cells from passive storage vats into dynamic regulators of metabolism.

Coleman arrived at work by 7 a.m. each morning and was sure to be home by 5.30 p.m. for family dinner. He was a keen hiker and sailor, and hunted for birds with a series of Brittany spaniels. Coleman was happiest working alone, assisted by two technicians at most. Over time, he concluded that his solitary style of research had become archaic. Satisfied with his achievements, he retired at age 62 rather than trying to reinvent his laboratory.

Coleman spent the following decade travelling the world with Beverly, always returning to their house by the sea in rural Maine. They were deeply committed to protecting lands and forests. He bequeathed their 20-hectare woodland and most of his assets to promote conservation.

Coleman felt that he had a great life despite, in his words, having endured two terrible tragedies: the loss of his 11-year-old son John to bone cancer and of Beverly to Alzheimer's disease in 2006. Tellingly, he did not list as a tragedy his own battle with an aggressive carcinoma of the face. Stoking his wood stove each morning, he found things to look forward to despite constant pain and eventual blindness. He also enjoyed the long-delayed recognition he so richly deserved, travelling the world with his sons and friends to receive awards.

Coleman walked with a brisk gait that gave the impression that he was headed somewhere. He spoke the same way, always amiably, but with purpose. When he did something, he did it meticulously and with conviction. Content to go where the data led him, he followed a path that helped to illuminate the molecular machinery that underlies a major issue in public health. ■

Jeffrey Friedman *is professor of molecular genetics at Rockefeller University in New York City. He was a co-recipient with Douglas Coleman of the 2010 Albert Lasker Basic Medical Research Award for the discovery of leptin.*
e-mail: friedj@mail.rockefeller.edu

CLIMATE SCIENCE

A sink down under

The finding that semi-arid ecosystems in the Southern Hemisphere may be largely responsible for changes in global concentrations of atmospheric carbon dioxide has repercussions for future levels of this greenhouse gas. SEE LETTER P.600

DANIEL B. METCALFE

The steady increase in atmospheric carbon dioxide, which is the main driver of global climate change, has fortunately been slowed by a simultaneous rise in CO_2 absorption by land plants, termed the land CO_2 sink[1]. However, exactly where and why this land sink is occurring has been uncertain. The main player had been thought to be tropical rainforest[2]. But on page 600 of this issue, Poulter *et al.*[3] draw together several lines of evidence to show that this situation may be changing, with vegetation across semi-arid ecosystems in the Southern Hemisphere taking centre stage as a driver of variations in atmospheric CO_2 levels.

Of the roughly 10 billion tonnes of carbon emitted each year from human activity, only around half remains in the atmosphere, with the rest being absorbed by the oceans and by plants on land[1]. This CO_2 sink has been growing steadily, but the situation could change as shifts in climate and human land use intensify. One warning sign of potential change is that the land sink seems to be highly sensitive to variations in temperature and rainfall over yearly timescales[4]. Advances in our understanding of the mechanisms behind this short-term variation in CO_2, such as those reported by Poulter and colleagues, provide valuable clues about the probable longer-term trend in atmospheric CO_2 over the coming century.

Unsurprisingly, given the inherent complexity of the atmosphere–land–ocean system, isolating exactly which components of the global carbon cycle are affected by yearly climate variation has proved difficult. However, large-scale assessments of carbon budgets have indicated that the culprit is somewhere on land[5], and simulations went one step further by pinpointing tropical forests as the prime suspect[2,4,6]. So far, so simple — until 2011, when scientists estimated an extraordinarily high value for the land sink which seemed to be linked not to tropical forest, but to semi-arid ecosystems in the Southern Hemisphere[7].

Enter Poulter and co-workers, who provide the most detailed account so far of the spatial patterns, and underlying mechanisms, of the record-high land sink of 2011, and more generally, of yearly variation in the land sink over the past few decades. Their results challenge the current consensus about what regulates atmospheric CO_2 from year to year, and will prove invaluable as societies struggle to predict, and adapt to, changes in a world in which both atmospheric and ecological systems are moving into unfamiliar territory.

JOHN WHITE PHOTOS/GETTY

Figure 1 | Dry drivers. Poulter *et al.*[3] report that semi-arid ecosystems in the Southern Hemisphere, such as in central Australia (pictured), drive yearly changes in global concentrations of atmospheric carbon dioxide.

Specifically, the authors find remarkable agreement in the magnitude of the 2011 land sink estimated from three contrasting approaches: a detailed ecosystem model simulating land–atmosphere interactions; an 'atmospheric inversion', which uses a gas-transport model to convert measurements of atmospheric CO_2 concentrations to estimates of surface CO_2 emissions; and a large-scale carbon budget, estimating land CO_2 absorption as the difference between human CO_2 emissions and the sum of the observed atmospheric CO_2 build-up and the relatively well-constrained oceanic CO_2 sink. Both the model and the inversion suggest that semi-arid ecosystems in the Southern Hemisphere (Fig. 1) are becoming increasingly valuable land CO_2 sinks — a conclusion supported by satellite data, which yield a visual record of plant growth and photosynthesis over the past 30 years.

A switch in the type of ecosystem that controls atmospheric CO_2 has major implications for the overall rate and pattern of climate change. First, different vegetation types may have different responses and sensitivities to disturbance. For example, Poulter *et al.* show that increased growth of vegetation in semi-arid ecosystems is linked not only to generally increasing levels of rainfall across these regions, but also to increasing sensitivity of the vegetation to rainfall. Worryingly, the authors also show that these links between rainfall and the land CO_2 sink in semi-arid ecosystems are currently missing from many major climate models.

Second, different types of ecosystem lock away absorbed CO_2 for different lengths of time. Rainforests store much of their carbon in dense hardwoods, which may take many centuries to die and rot[8,9], whereas much of the CO_2 absorbed across semi-arid regions is converted into relatively short-lived grasses and shrubs[10,11]. Increased effects of these more volatile ecosystems on global climate could lead to greater variability in global atmospheric CO_2 levels in the future, something that recent observations seem to confirm.

Although Poulter and colleagues use the best available models to support their conclusions, models are only as good as the data with which

they are calibrated, and little information exists about vegetation in semi-arid ecosystems compared with other regions. Similarly, the pathways for CO_2 once it has been absorbed into semi-arid vegetation remain poorly understood[10,11], so there are few solid data from which to assess the stability of the CO_2 sink in such ecosystems. More broadly, semi-arid systems are vulnerable to a range of factors that are difficult to model, such as overgrazing, fire, flooding and chronic soil erosion[10,11], many of which are linked to human activity. These processes must somehow be accounted for, both in models and in policies for land use and conservation, if the invaluable function of semi-arid ecosystems as a global CO_2 sink is to be managed and maintained.

Nevertheless, Poulter *et al.* make a key contribution in highlighting the crucial, and hitherto often overlooked, role of such ecosystems in the global carbon cycle, and in identifying several important processes, which should markedly improve our understanding of future atmospheric CO_2 levels. Let us hope that their research stimulates more work on the ground to better understand and manage these fragile but essential ecosystems. ■

Daniel B. Metcalfe *is in the Department of Physical Geography and Ecosystem Science, Lund University, 223 62 Lund, Sweden.*
e-mail: dbmetcalfe@gmail.com

1. Ballantyne, A. P., Alden, C. B., Miller, J. B., Tans, P. P. & White, J. W. C. *Nature* **488,** 70–72 (2012).
2. Cox, P. M. *et al. Nature* **494,** 341–344 (2013).
3. Poulter, B. *et al. Nature* **509,** 600–603 (2014).
4. Zeng, N., Mariotti, A. & Wetzel, P. *Glob. Biogeochem. Cycles* http://dx.doi.org/10.1029/2004GB002273 (2005).
5. Bousquet, P. *et al. Science* **290,** 1342–1346 (2000).
6. Sitch, S. *et al. Glob. Change Biol.* **14,** 2015–2039 (2008).
7. Bastos, A., Running, S. W., Gouveia, C. & Trigo, R. M. *J. Geophys. Res.* **118,** 1247–1255 (2013).
8. Chambers, J. Q., Higuchi, N. & Schimel, J. P. *Nature* **391,** 135–136 (1998).
9. Chambers, J. Q., Higuchi, N., Schimel, J. P., Ferreira, L. V. & Melack, J. M. *Oecologia* **122,** 380–388 (2000).
10. Scurlock, J. M. O. & Hall, D. O. *Glob. Change Biol.* **4,** 229–233 (1998).
11. Serrano-Ortiz, P., Sánchez-Cañete, E. P. & Oyonarte, C. in *Recarbonization of the Biosphere: Ecosystems and the Global Carbon Cycle* (eds Lal, R., Lorenz, K., Hüttl, R. F., Schneider, B. U. & von Braun, J.) Ch. 15, 347–368 (Springer, 2012).

This article was published online on 21 May 2014.

MICROBIOLOGY

Barriers to the spread of resistance

Despite identifying abundant genes capable of conferring antibiotic resistance in soil microorganisms, a study finds that few are shared by human pathogens and that there is little transfer of the genes within the soil communities. SEE LETTER P.612

MORTEN O. A. SOMMER

Antibiotic resistance is complicating the treatment of many bacterial infections, leading to increasing mortality and health-care costs across the globe. Pathogens frequently acquire resistance to antibiotics from other sources, and soil-dwelling bacteria are considered an important reservoir of resistance genes. In this issue, Forsberg *et al.*[1] (page 612) identify nearly 3,000 genes conferring antibiotic resistance from the soil, and find that only a minute fraction of these genes is shared with human pathogens. Furthermore, their data suggest that it is the mobilization of and selection for such genes, rather than their supply, that limits their transfer among soil bacteria and with other bacteria, including human pathogens.

The genome sequencing of thousands of bacterial pathogens has shown that antibiotic-resistance genes are often acquired by pathogens from other sources through horizontal gene transfer[2]. This process allows even distantly related bacteria to transfer genes through the action of viruses (bacteriophages) or conjugative plasmids (small DNA molecules separate from chromosomal DNA that have the machinery to facilitate transfer between bacteria), or through uptake of free-floating DNA in the environment. All of these mechanisms are known to contribute to the emergence of antibiotic resistance in human pathogens, yet the origin of most clinically relevant resistance genes remains elusive. Accordingly, research efforts are directed at addressing this question through the elucidation of antibiotic-resistance genes harboured by different microbial communities[3].

Several soil bacteria can produce antibiotics, and these bacteria also have the necessary genes to confer immunity against the toxins they produce. Accordingly, it has been proposed[2] that these antibiotic producers could be an origin of antibiotic-resistance genes. Subsequent research has also pointed to other soil microbes that might serve as sources of resistance genes, including bacteria that can subsist on antibiotics[4]. Yet, in spite of substantial research in this area, only a few studies have demonstrated a link between resistance genes in the soil and resistance genes in human pathogens[5].

Forsberg and colleagues deployed functional metagenomics to study soil microbial communities and characterize genes that confer antibiotic resistance on a non-resistant strain of the bacterium *Escherichia coli*. They identified nearly 3,000 resistance-conferring genes,

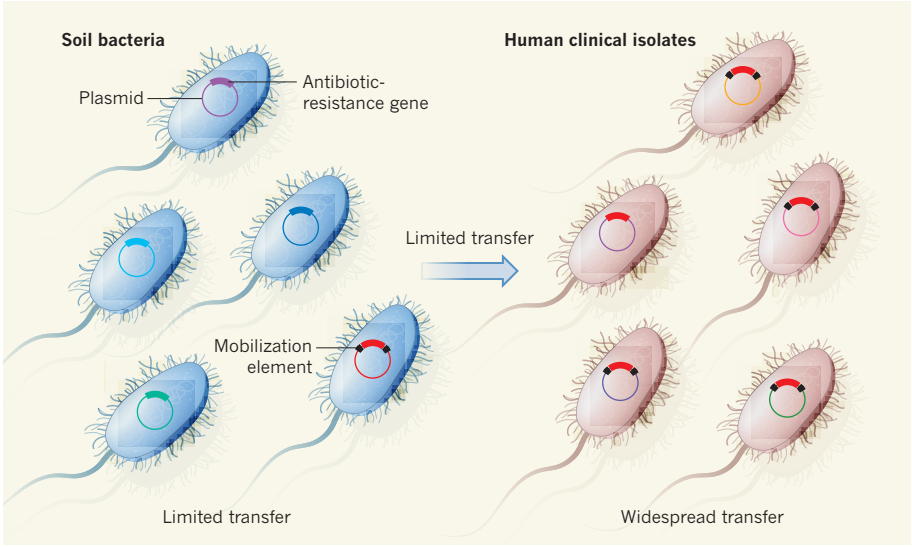


Figure 1 | Limited transfer. Forsberg and colleagues' metagenomic analysis of soil microorganisms[1] identified around 3,000 genes capable of conferring antibiotic resistance, but found that less than 0.1% of these genes have been identified as associated with antibiotic resistance in bacteria isolated from human patients. They also show that fewer resistance genes in the soil bacteria are flanked by mobilization elements than those in pathogens, suggesting that there is limited transfer of these genes within the soil community and from the soil to other bacteria.

which is a number comparable to all currently known antibiotic-resistance genes[6]. Thus, this study shows clearly that an extraordinary diversity of antibiotic-resistance genes exists in nature, as suggested by previous analysis of soil microbial communities[7].

An earlier study[5] from the group presenting the current paper reported the first case of the transfer of several drug-resistance genes between innocuous soil bacteria and human pathogens, highlighting that transfer of genes between such bacteria is possible. But that study included bacterial-enrichment steps that prevent quantification of the extent of such transfer. In the present study, the authors used a method that did not require enrichment and which allowed them to quantify the extent of antibiotic-resistance genes that are shared between soil bacteria and previously characterized bacteria.

They found that only around 0.1% of the identified resistance genes from soil are highly similar (greater than 99% nucleotide identity) to previously detected resistance genes, indicating that there is only limited overlap between the resistance genes of soil bacteria and other bacteria, including those that cause infections in humans (Fig. 1). Although this low overlap does not exclude the possibility of soil bacteria acting as an origin of antibiotic-resistance genes that cause clinical problems, it does demonstrate that only a minute fraction of resistance genes from soil bacteria have been transferred to human pathogens.

Forsberg *et al.* also investigated whether the limited overlap might result from limited transfer of antibiotic-resistance genes within the soil microbial community. If this is the case, specific resistance genes should be stably associated with specific phylogenetic divisions. The authors show that this is correct and conclude that the resistance-gene pool of different soil communities is closely linked to the phylogenetic architecture of those communities. The authors were not able to resolve the phylogenetic architecture beyond the phylum level, and so horizontal transfer of genes within a specific phylum cannot entirely be ruled out. However, they show that soil bacteria, in contrast to human pathogens, have a much lower number of mobilization elements flanking their resistance genes, which supports their hypothesis of limited transfer of resistance genes between soil bacteria. These results are consistent with the hypothesis that there is limited selection for antibiotic resistance within the soil microbiota compared to the selection for antibiotic resistance in human pathogens.

These findings fuel the ongoing question of what is the function of antibiotic-resistance genes in their natural hosts. For instance, the MFS transporter proteins identified by the authors as conferring resistance to a wide range of antibiotic classes may not actually function as antibiotic-resistance proteins in their hosts, but rather in different processes, such as the transport of other small molecules that may be more abundant than antibiotics in the soil. Similarly, the identified β-lactamase enzymes might serve as cell-wall remodelling enzymes in their natural hosts. The apparent paucity of mobilization elements flanking these genes would suggest that selection for and transfer of resistance functions in the soil is not as strong as in other environments.

Irrespective of the function of these genes in their natural hosts, Forsberg and colleagues' study demonstrates that the soil microbiota harbours an extraordinary diversity of genes that have the potential to confer antibiotic resistance in human pathogens such as *E. coli*. Their findings also suggest that it may not be the availability of genes encoding proteins capable of conferring antibiotic resistance that limits the spread of resistance, but rather the mobilization and transfer of these genes. Functional metagenomic studies of soils that have been exposed to inhibitory concentrations of antibiotics should be performed to test whether this increases the extent of resistance-gene mobilization. ■

Morten O. A. Sommer *is in the Department of Systems Biology and the Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kongens Lyngby, Denmark.*
e-mail: msom@bio.dtu.dk

1. Forsberg, K. J. *et al. Nature* **509,** 612–616 (2014).
2. Davies, J. & Davies, D. *Microbiol. Mol. Biol. Rev.* **74,** 417–433 (2010).
3. Allen, H. K. *et al. Nature Rev. Microbiol.* **8,** 251–259 (2010).
4. Dantas, G., Sommer, M. O. A., Oluwasegun, R. D. & Church, G. M. *Science* **320,** 100–103 (2008).
5. Forsberg, K. J. *et al. Science* **337,** 1107–1111 (2012).
6. McArthur, A. G. *et al. Antimicrob. Agents Chemother.* **57,** 3348–3357 (2013).
7. Riesenfeld, C. S., Goodman, R. M. & Handelsman, J. *Environ. Microbiol.* **6,** 981–989 (2004).

This article was published online on 21 May 2014.

MATERIALS SCIENCE

Energy storage wrapped up

Cables and wires are used to conduct electricity, but can they also store energy? The answer is a resounding 'yes', if they are encased by a supercapacitor device — a finding that might open up many applications.

YURY GOGOTSI

Electrical cables entangle the world, supplying electricity to buildings, machines and electronic devices. The systems currently used to store electrical energy are separate from the cables, and are bulky contraptions often consisting of assemblies of 'supercapacitor' devices. Reporting in *Advanced Materials*, Yu and Thomas[1] describe coaxial cables consisting of a copper core surrounded by a supercapacitor sheath, which can both transmit and store electricity.

Energy storage in supercapacitors can involve two mechanisms[2]: the formation of a double layer of ions adsorbed on oppositely charged electrode surfaces; and pseudocapacitance, in which fast electrochemical reactions occur at the surface of an electrochemically active material, such as manganese dioxide. Because pseudocapacitance occurs on a large electrode surface, it always takes place alongside double-layer capacitance.

In supercapacitors, charge is stored only at surfaces, and so — unlike in batteries — its availability is not limited by diffusion processes, allowing high power to be achieved[3]. Similarly, because charging and discharging do not involve a bulk-phase transformation, as they do in batteries, supercapacitors are much more reversible (less energy is lost during a charge–discharge cycle) and have a longer cycle life[2] (up to a million charge–discharge cycles). These properties are desirable for energy-storing cables.

To add capacitive storage to conventional wires, Yu and Thomas effectively wrapped a supercapacitor around a core conductor wire (Fig. 1). They began by growing nanowires of insulating copper oxide perpendicular to the surface of a copper wire, and then coated these nanowires with a gold–palladium alloy, which acts as a current collector for the supercapacitor. An electrochemically active coating of manganese oxide was then deposited on top of the alloy. The resulting brush-like architecture leads to a 100-fold increase in surface area compared with the bare copper wire; a large surface area is crucial for capacitive energy storage. The nanowires serve as a sheath covering the copper wire, and form the first electrode of the supercapacitor.

To construct the rest of the device, the authors coated this electrode with a solid electrolyte (a material that conducts ions, but not electrons, and which electrically connects

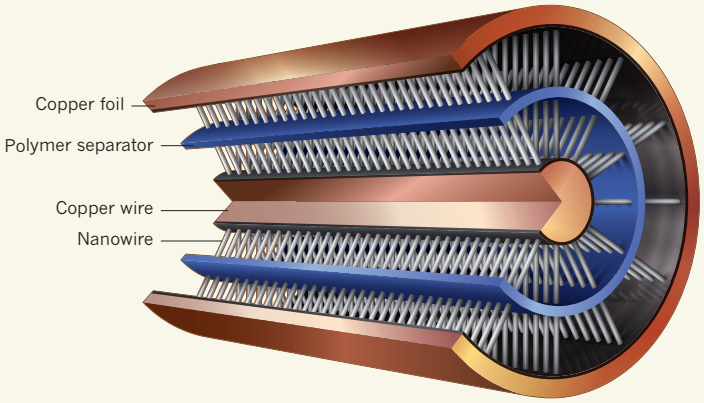


Figure 1 | Energy-storing cables. Yu and Thomas[1] report cables that conduct electricity through a central wire, but that also store electrical energy in a coaxial supercapacitor. In the authors' design, electrochemically active nanowires project from a copper wire coated with copper oxide (black). The nanowire layer forms the first electrode of the supercapacitor device, and is wrapped in a porous polymer separator. A second electrode — consisting of nanowires as described above, attached to a copper foil coated internally with copper oxide — surrounds the polymer separator. Both electrodes are coated in a gel electrolyte (not shown) before assembly, to ensure that they are electrically connected to each other. (Figure adapted from ref. 1.)

the supercapacitor's electrodes) and a porous polymer separator, and then assembled another tubular electrode around that. The second electrode was made in the same way as the first, except that nanowires were formed on a copper foil, rather than a wire. The resulting material was also precoated with a solid electrolyte before its assembly around the first electrode. The coaxial geometry of the resulting device combines all of the components of a supercapacitor into a single cable and ensures good electrical contact between the electrodes and current collectors. The researchers observed that the core wire can carry current independently of energy storage, at least when low-voltage direct current is transmitted through it.

Yu and Thomas report a high capacitance per unit mass of active material for their device — much higher than that of other pseudocapacitors based on manganese dioxide. However, this was the case only for a small loading of 0.5 milligrams of oxide per square centimetre. The values reported for devices with larger loadings were much smaller, which limits the amount of energy that could be stored in these coaxial supercapacitors. Increasing the amount of stored energy will be essential for practical applications.

Energy-storing cables must be flexible, strong, wear-resistant and well insulated, and no electrode fracturing or short-circuiting should occur on bending. The researchers bent their device at different angles up to 100 times and found that the capacitance was generally stable, although folding it 100 times at 180° led to a loss of about 7% of the initial capacitance. Furthermore, only 1% of capacitance was lost after 5,000 charge–discharge cycles without folding.

Single supercapacitor cells usually produce 1 to 3 volts, and generate direct current. Devices with multiple cells connected in sequence can achieve the 12–24 V needed for applications such as small electronic devices, humanoid robots and automotive electronics, all of which use direct current. Household applications, however, use higher voltages (110 V in the United States and 220 V in Europe) and alternating current. Supercapacitors are impractical for these applications because of the need for conversion from direct to alternating current, and because about 100 cells would need to be connected in sequence, which would lead to energy losses associated with electrical resistance. Losses would also occur because of the high electrical field that is generated when alternating current is transmitted through a wire.

More-realistic applications of Yu and Thomas's devices will be in cables connecting generators of renewable energy, to level up power when individual generators are not producing as much energy as their neighbours — for example, when a cloud passes over a solar-energy farm. The devices might also be useful for storing the electrical energy produced by solar panels or wind-power generators[4]. Furthermore, coaxial cables that store energy (using either supercapacitors or coaxial batteries[5,6]) might help to miniaturize electronic devices by decreasing the size of the bulky batteries that are currently used. Alternatively, capacitive storage could be used to increase the lifetime of batteries in electronic equipment, and the time between charges. However, the adoption of energy-storing cables as replacements for stand-alone supercapacitors or batteries will depend on whether low-cost materials such as carbon can be used, and on the development of simple manufacturing processes.

Finally, there is a tremendous and rapidly growing interest in energy-storage devices and systems that are flexible, wearable and that can be integrated into textiles[7,8]. Lightweight, flexible coaxial cables could be used in knitted textiles. If the cable is too thick or

50 Years Ago

'Science in a state of siege' — At that same meeting I was trying to explain the work on polyelectrolytes which had started in the Weizmann Institute. I was interrupted by a woman who asked, "Polly who?" I said, 'Polly nobody" and went on to explain the polyelectrolytic phenomenon … Briefly and over-simply polyelectrolytes are long-chain macromolecules … which have become coiled up and can uncoil … The homeliest example I can think of is a table jelly; you put crystals, or a gelatinous slab, in water and the polyelectrolyte molecules uncoil and expand to fill the jelly-mould. So housewives have been experimenting with polyelectrolytes for a very long time. Prof. Richie Calder

From *Nature* 30 May 1964

100 Years Ago

Sounds and Signs: a Criticism of the Alphabet with Suggestions for Reform by Archer Wilde — The main object of this book is to advocate modifications in our present alphabet, so as to make it suitable for representing English sounds … the suggested alphabet is portrayed; the capitals are practically identical with the small letters, but slightly more ornate. A characteristic is that no letter projects above or below the line; nor are parts of each letter thicker or thinner than others; the character is what is termed "Doric." … In the example … of printing in Doric capitals, the effect is to dazzle the eyes; it is not easy reading. The author is not sanguine as to the adoption of his scheme; but he opens the interesting question whether if our alphabet is to be modified, convenience is to be increased by carefully choosing the form of the letters.

From *Nature* 28 May 1914

inflexible for knitting, it could still be embedded in textiles to provide power and transmit signals to and from sensor arrays implanted into clothes[7,8]. Indeed, coaxial-fibre supercapacitors for textile electronics have already been reported[9], in which a core electrode 230 micrometres in diameter was made from bundles of carbon microfibres coated with multi-walled carbon nanotubes, with carbon-nanofibre paper acting as the outer electrode. All-carbon electrodes are light and stable, and could potentially be wrapped around copper wires (including those bearing copper oxide nanowires) to form coaxial devices similar to those of Yu and Thomas. ■

Yury Gogotsi *is at the A. J. Drexel Nanomaterials Institute, Department of Materials Science and Engineering, Drexel University, Philadelphia, Pennsylvania 19104, USA.*
e-mail: gogotsi@drexel.edu

1. Yu, Z. & Thomas, J. *Adv. Mater.* http://dx.doi.org/10.1002/adma.201400440 (2014).
2. Conway, B. E. *Electrochemical Supercapacitors: Scientific Fundamentals and Technological Applications* (Kluwer, 1999).
3. Simon, P., Gogotsi, Y. & Dunn, B. *Science* **343,** 1210–1211 (2014).
4. Lindley, D. *Nature* **463,** 18–20 (2010).
5. Christodoulou, L. & Venables, J. D. *JOM* **55** (12), 39–45 (2003).
6. Kwon, Y. H. *et al. Adv. Mater.* **24,** 5192–5197 (2012).
7. Jost, K., Anasori, B., Beidaghi, M., Dion, G. & Gogotsi, Y. in '2013 visualization challenge' *Science* **343,** 600–610 (2014).
8. Jost, K., Dion, G. & Gogotsi, Y. *J. Mater. Chem. A* http://dx.doi.org/10.1039/c4ta00203b (2014).
9. Le, V. T. *et al. ACS Nano* **7,** 5940–5947 (2013).

DEVELOPMENTAL GENETICS

Female silkworms have the sex factor

Sex determination in the silkworm *Bombyx mori* has been found to depend on the presence or absence of a small RNA. This is thought to be the first example of a molecule other than a protein mediating this process. SEE LETTER P.633

FRANTIŠEK MAREC

Over the past 80 years, a growing body of evidence[1,2] has suggested that, in the silkworm *Bombyx mori,* sex is determined by the presence or absence of a maternally inherited sex chromosome, known as W. But despite years of searching for a protein-coding gene on W that could be the primary trigger for development as a female, the molecular mechanism of sex determination in the silkworm has remained unknown[3]. On page 633 of this issue, Kiuchi *et al.*[4] report the discovery of a surprising, RNA-mediated pathway of sex determination in *B. mori*.

In many insects, the development of sexual characteristics involves an evolutionarily conserved mechanism: sex-specific regulation of the gene *doublesex* (ref. 5). In *B. mori*, as in the fruit fly *Drosophila melanogaster*, transcription of *doublesex* produces different forms of messenger RNA in males and females[6], a process called alternative splicing. The different mRNA transcripts then turn on sex-specific gene expression, giving rise to male or female traits[5]. But what is the sex-determining factor that controls splicing of *Bombyx mori doublesex* (*Bmdsx*)?

In silkworms, as in most members of Lepidoptera (moths and butterflies), sex is dependent on 'WZ' sex chromosomes, in contrast to the XY system found in many other species, including mammals. In WZ systems, an embryo with one W and one Z chromosome develops as a female, and an embryo with two Z chromosomes develops as a male. The W chromosome of *B. mori* lacks protein-coding genes and consists mainly of transposons — mobile genetic elements that move around within the genome and can thus cause mutations. The only transcripts produced from the W chromosome are PIWI-interacting RNAs (piRNAs), small RNAs that inhibit transposons in animals' gonads by interacting with PIWI proteins[7].

To identify potential sex determinants on W, Kiuchi *et al.* performed in-depth sequencing of the RNA transcripts expressed in female and male *B. mori* embryos. They identified one transcript that is expressed in females at all stages of development, but never in males. This transcript is a precursor of a female-specific piRNA, which the researchers name *Fem* piRNA.

The authors report that inhibiting the expression of *Fem* piRNA in female embryos changes the splicing pattern of *Bmdsx* mRNA from the female to the male form, suggesting that this piRNA is required for female sex determination. Furthermore, they demonstrate that *Fem* piRNA targets and cleaves an mRNA molecule transcribed from a gene on the Z chromosome, *Masculinizer* (*Masc*), which encodes the zinc-finger protein masculinizer (Masc).

Kiuchi and colleagues find that, in the absence of *Fem* piRNA (in embryos lacking the W chromosome), Masc promotes male-specific splicing of *Bmdsx*, resulting in male development. *Fem* piRNA inhibits male development in WZ embryos by down-regulating the level of Masc (Fig. 1). Female-specific *Bmdsx* splicing then occurs by default, ensuring female development. Thus, the W-encoded *Fem* piRNA is the long-sought primary trigger of female development.

The detailed molecular mechanism revealed by this work may shed light on sex determination in most lepidopterans. Exceptions to the WZ system include 'primitive' moths such as the Micropterigidae and Hepialidae (which arose before the evolution of W and so lack this chromosome and have a Z/ZZ system like their closest relatives, caddisflies) and some 'advanced' species, such as the Eri silkworm *Samia cynthia ricini*, in which the W chromosome has been lost[8].

Because W chromosomes are transmitted only through females, they cannot undergo recombination — a mechanism that generates genetic variation in other chromosome pairs, and that also prevents mutations that

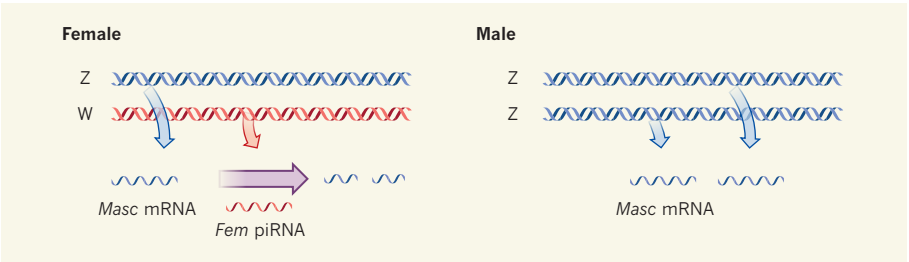


Figure 1 | Sex determination in silkworms. In most moths and butterflies, sex is dependent on whether or not the embryo has a W chromosome — embryos with one W and one Z chromosome develop as females, whereas embryos with two Z chromosomes become male. Kiuchi *et al.*[4] report that a PIWI-interacting RNA (*Fem* piRNA) molecule transcribed from the W chromosome is responsible for sex determination in the silkworm *Bombyx mori*. The authors find that *Fem* piRNA cleaves *Masc* messenger RNA, which is transcribed from the Z chromosome. In ZZ embryos, expression of *Masc* promotes the expression of 'male' genes. In WZ embryos, *Fem* piRNA decreases levels of *Masc*, leading to female development by default.

negatively affect gene function from becoming fixed in the population. As a result, W chromosomes in different species have common features, including extensive genetic erosion, accumulation of repetitive DNA sequences and uniformity in DNA composition[8]. Why female genomes in most lepidopterans still have these chromosomes, despite their apparent uselessness and lack of protein-coding genes, might therefore seem unclear. The piRNA-driven pathway discovered by Kiuchi *et al.*[4] could provide a rationale for their existence.

How, then, is sex determined in lepidopterans that lack a W chromosome? Kiuchi and colleagues point out that the occurrence of genes equivalent to *Masc* in several other lepidopteran species may indicate an evolutionarily conserved role for *Masc* in sex determination. In such cases, a Z-counting mechanism might control sex, with a male-promoting gene such as *Masc* on the Z chromosome acting against a female-promoting gene on another chromosome — expression of *Masc* from two Z chromosomes being required to tip the balance in favour of male development. Such an ancestral Z-counting mechanism might have been re-established in species that lost the W chromosome at a later date.

The only weakness in Kiuchi and co-workers' data is the absence of a conclusive assay showing that inhibition of either *Fem* piRNA or *Masc* expression can cause a change of sex from that expected from an embryo's chromosomal make-up. An inhibitor of *Fem* piRNA injected into WZ silkworm embryos proved to be insufficiently effective to cause male development, and genetic inhibition of *Masc* caused ZZ embryos to die before hatching. Nevertheless, when conducting RNA sequence analysis of the *Masc*-inhibited ZZ embryos before death, the authors made another observation — the Masc protein is required for dosage compensation, the process by which the expression of genes on the Z chromosome is downregulated in ZZ embryos to lower the genes' transcription to the same level as in ZW embryos.

Dosage compensation has not previously been reported in lepidopterans. Kiuchi and colleagues' evidence that this occurs in *B. mori* contradicts results showing[8] that male lepidopterans tolerate a double dose of Z-linked gene transcripts compared with females, but is consistent with data[9] on global gene expression in the silkworm indicating that gene expression from the Z chromosome occurs at the same level in males and females. As the authors note, the lethal effect of *Masc* inhibition is similar to the 'male-killing action' of the bacterium *Wolbachia* in several lepidopteran species: infection by *Wolbachia* kills male embryos[10]. It seems that this, too, may result from a failure of dosage compensation.

This is thought to be the first example of a sex-determining pathway controlled by the presence or absence of a piRNA. Whether this pathway is evolutionarily conserved in lepidopterans will certainly be the subject of future research. Kiuchi and co-workers' findings also pave the way to understanding the molecular mechanism by which *Wolbachia* and other endosymbiotic bacteria manipulate the sex of their host. Finally, the results open up options for artificial sex control in silkworms (for example, to promote the development of males, which produce more silk than do females[2]) and possible ways to control lepidopteran pests. ■

František Marec *is at the Institute of Entomology, Biology Centre ASCR, České Budějovice 370 05, Czech Republic.*
e-mail: marec@entu.cas.cz

1. Hasimoto, H. *Jpn. J. Genet.* **8,** 245–247 (1932–33).
2. Tazima, Y. *The Genetics of the Silkworm* (Academic, 1964).
3. Suzuki, M. T. *J. Genet.* **89,** 357–363 (2010).
4. Kiuchi, T. *et al. Nature* **509,** 633–636 (2014).
5. Cho, S., Huang, Z. Y. & Zhang, J. *Genetics* **177,** 1733–1741 (2007).
6. Suzuki, M. G., Funaguma, S., Kanda, T., Tamura, T. & Shimada, T. *Evol. Dev.* **7,** 58–68 (2005).
7. Kawaoka, S. *et al. RNA* **17,** 2144–2151 (2011).
8. Traut, W., Sahara, K. & Marec, F. *Sexual Dev.* **1,** 332–346 (2007).
9. Walters, J. R. & Hardcastle, T. J. *Genome Biol. Evol.* **3,** 491–504 (2011).
10. Kageyama, D. & Traut, W. *Proc. R. Soc. B* **271,** 251–258 (2004).

This article was published online on 14 May 2014.

PRECISION MEASUREMENT

The magnetic proton

A record measurement of the proton's magnetism has been achieved by confining a single proton in a device called a double Penning trap. The result opens the way to exploring one of nature's fundamental symmetries. **SEE LETTER P.596**

V. ALAN KOSTELECKÝ

The achievement of an unprecedented degree of precision in the measurement of the proton's magnetic moment is described by Mooser *et al.*[1] in this issue (page 596). This impressive feat, obtained by trapping and studying one proton at a time, sets the stage for a new test of a profound symmetry of nature and of Einstein's relativity.

Nature's secrets are encoded in fundamental laws of physics that govern the properties and behaviours of elementary particles. Ordinary matter consists of atoms formed from electrons, protons and neutrons. Unlike the electron, for which no evidence of substructure has been found, the proton and the neutron are each composed of elementary particles called quarks and gluons. It is impossible to isolate quarks or gluons because they interact too strongly with one another, so one way to determine their properties indirectly is to measure the properties of the proton.

Experimentally, the proton is found to be a stable particle with mass and electric charge. It also behaves like a tiny magnet, the strength of which is called the proton magnetic moment. Before Mooser and colleagues' work, the most precise measurement of the proton magnetic moment dated from more than 40 years ago[2]. This early study achieved a precision of about 10 parts per billion (p.p.b.) by means of spectroscopic studies of atomic hydrogen and input from separate theoretical calculations.

The trapping, storage and study of a single particle is an alternative and direct approach to precision measurements. For example, a device called a Penning trap uses a combination of magnetic and electric fields to keep a charged particle confined in a small region. This method has been used to measure the magnetic moment of the electron at an impressive precision of 0.0003 p.p.b. (ref. 3).

The application of this technique to the proton is much more challenging, because the proton magnetic moment is about 658 times smaller than that of the electron. It involves a magnetic field with a larger spatial inhomogeneity that, in turn, makes obtaining a precision measurement more difficult. Nonetheless, measurements of the proton magnetic moment using a Penning trap to confine and probe a single trapped proton have reached a sensitivity of about 2,500 p.p.b. (refs 4, 5).

To achieve the most precise measurement so far, Mooser *et al.* used a double Penning trap, in which two Penning traps are conjoined such that their midpoints are separated by about 5 centimetres. One trap has a magnetic field with a large spatial inhomogeneity and is used to analyse the quantum state of the proton; the other has a homogeneous field and is used to make precision measurements. In the experiment, the proton is shuttled back and forth between the traps and its quantum state manipulated until a measurement can be made, a process that takes about 2 hours. By analysing data taken over the course of 4 months, the authors were able to attain a record precision of 3.3 p.p.b. for the proton magnetic moment.

A primary motivation for making precision measurements of the proton magnetic moment is the prospect of comparing it to the magnetic moment of the proton's antiparticle, the antiproton. Antiprotons are observed experimentally to have the same mass as protons and an

opposite charge, in accordance with a profound theoretical result from multi-particle quantum physics called the CPT theorem. The essence of this theorem is that performing a charge conjugation (C, which interchanges particles and antiparticles), a parity inversion (P, which involves a mirror reflection and a rotation) and a time reversal (T, which changes the direction of the flow of time) leaves physical laws unchanged, and so the combined CPT transformation represents a symmetry of nature. The theorem holds for realistic multi-particle quantum theories only if Einstein's relativity is exactly valid, so tiny deviations from CPT symmetry would be accompanied by tiny violations in the laws of relativity[6]. One consequence of the theorem is that the proton and antiproton magnetic moments must be equal in magnitude, so comparing these two quantities experimentally offers a sharp test of CPT symmetry, and therefore an opportunity to search for the tiny relativity violations predicted in some theories of nature.

In practice, any difference in the observed proton and antiproton magnetic moments would emerge from shifts in their quantum energies and would involve two factors, one constant and the other varying with time owing to the motion of the laboratory as Earth rotates on its axis and revolves around the Sun[7]. Disentangling these two effects requires a large data set, and a detailed experimental study with protons and antiprotons has not yet been published. Pioneering experiments of this type have used electrons and positrons (antielectrons) in Penning traps to constrain both constant[8] and time-varying[9] energy shifts violating the CPT theorem to about 2 parts in 10^{21} of the electron's rest energy (0.511 megaelectronvolts). Mooser and colleagues' new techniques offer one promising route by which to extend these tests to protons and antiprotons.

The prospects for future improvements in the measurement precision of the proton magnetic moment are excellent. For the double Penning trap, further reducing the field inhomogeneity and sharpening the experimental procedure are expected to increase precision by a factor of ten[1]. A different and ambitious scheme now under development involves a precision array with two Penning traps, one containing a proton or antiproton, and the other an atomic ion[10]. The ion would improve the control of the magnetic field and the measurement procedure, and the timescale required for a measurement would be reduced from about 2 hours to approximately a second. These and other future experiments on protons and antiprotons will stringently test and enhance our understanding of the fundamental laws of nature. ■

V. Alan Kostelecký *is in the Physics Department and the Center for Spacetime Symmetries, Indiana University, Bloomington, Indiana 47405, USA.*
e-mail: kostelec@indiana.edu

1. Mooser, A. *et al. Nature* **509,** 596–599 (2014).
2. Winkler, P. F., Kleppner, D., Myint, T. & Walther, F. G. *Phys. Rev. A* **5,** 83–114 (1972).
3. Hanneke, D., Fogwell, S. & Gabrielse, G. *Phys. Rev. Lett.* **100,** 120801 (2008).
4. Rodegheri, C. C. *et al. New J. Phys.* **14,** 063011 (2012).
5. DiSciacca, J. & Gabrielse, G. *Phys. Rev. Lett.* **108,** 153001 (2012).
6. Greenberg, O. W. *Phys. Rev. Lett.* **89,** 231602 (2002).
7. Bluhm, R., Kostelecký, V. A. & Russell, N. *Phys. Rev. D* **57,** 3932–3943 (1998).
8. Dehmelt, H., Mittleman, R., Van Dyck, R. S. & Schwinberg, P. *Phys. Rev. Lett.* **83,** 4694–4696 (1999).
9. Mittleman, R. K., Ioannou, I. I., Dehmelt, H. G. & Russell, N. *Phys. Rev. Lett.* **83,** 2116–2119 (1999).
10. Ospelkaus, C. *et al. Nature* **476,** 181–184 (2011).

CARDIOVASCULAR BIOLOGY

Switched at birth

Exposure to atmospheric oxygen in the days after a mammal's birth causes its heart muscle cells to stop proliferating. The finding may explain why zebrafish, which live in a hypoxic environment, can regenerate their hearts as adults.

KATHERINE E. YUTZEY

Cardiac muscle cells in the mature mammalian heart are notorious for their inability to proliferate in sufficient numbers to repair the heart after injury, or to prevent heart failure during the course of chronic disease[1]. But zebrafish and some amphibians can regenerate large parts of their hearts as adults, providing hope that regenerative mechanisms might be active in mammals under certain circumstances[2]. In fact, the hearts of newborn mice can regenerate after injury if it occurs within a few days of birth, but they lose this ability soon thereafter[3]. Writing in *Cell*, Puente *et al.*[4] show that the transition to an oxygen-rich environment after birth is what induces cardiac muscle cells to stop proliferating in mammals, but that this transition does not occur in zebrafish. These insights may have implications for efforts devoted to repairing, or ultimately regenerating, adult mammalian hearts.

The authors compared the proliferation and metabolism of heart muscle cells (cardiomyocytes) from neonatal mice and zebrafish. They found that exposure of newborn mice to a mildly hypoxic environment (15% oxygen, compared with the standard atmospheric level of 21%) delayed the withdrawal from the cell cycle that normally occurs in mammalian cardiomyocytes soon after birth. By contrast, hyperoxia (100% oxygen) accelerated this process, supporting the hypothesis that oxygen exposure at birth stops cardiomyocyte proliferation. On the basis of these results, the authors suggest that the hypoxic environments of the mouse fetus and adult zebrafish enhance cardiomyocyte proliferation and regenerative capacity, but that these characteristics are lost in mammals with exposure to atmospheric oxygen at birth.

In many ways — including its hypoxic environment, use of glucose as a source of energy and capacity of its cardiomyocytes to proliferate — the mouse fetus is more like a fish than the adult of its own species (Fig. 1). The neonatal period in mammals is a time of major trauma. All of an animal's organs receive multiple stimuli as a result of both exposure to atmospheric oxygen and the need to feed, move independently, secrete waste and

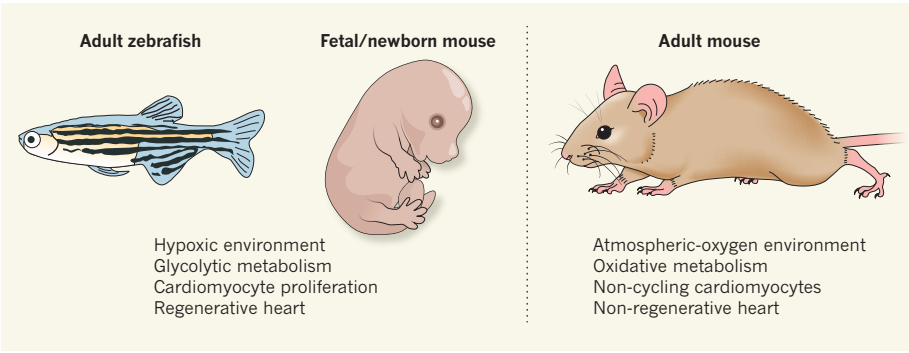


Figure 1 | Fish-like beginnings. The mammalian heart during fetal development and in the neonatal period has similarities to that of an adult zebrafish. Both exist in low-oxygen environments, and the heart muscle cells (cardiomyocytes) use non-oxidative, glucose-based (glycolytic) metabolism and are proliferative, such that they have the potential to regenerate a large part of their hearts after injury. By contrast, the adult mammalian heart exists in an oxygen-rich environment and generates energy by oxidative metabolism. Puente *et al.*[4] show that, in mice, the transition to oxygenated surroundings after birth leads to the withdrawal of cardiomyocytes from the cell cycle and the inability of the heart to regenerate.

regulate body temperature. Although Puente and colleagues single out changes in atmospheric oxygen as the cause of cardiomyocyte cell-cycle withdrawal, other factors — such as metabolic stress resulting from periods of starvation between birth and the initiation of nursing or intermittent feeding — could also contribute to cardiomyocyte maturation. However, the finding that oxygen-level manipulation is sufficient to affect the timing of cell-cycle withdrawal demonstrates that oxygen is a powerful stimulus in the developing heart.

Much of what is known of neonatal development is based on research in mice, which have a gestation period of 19–21 days. Humans are born with more-mature organs than mice, after a nine-month gestation period. The observation that neonatal mice can regenerate part of the heart after injury[3] has led to much speculation over potential treatment of human congenital heart defects. However, little is known about how human cardiomyocytes proliferate and mature during late fetal or neonatal development, and it is not known to what extent the development of the newborn mouse heart corresponds to that of a human heart. It is possible that exposure to atmospheric oxygen at birth has a comparable effect on human cardiomyocytes and their potential to regenerate, but studies of large animals with longer gestation periods than mice, or of non-human primates, will be necessary to determine whether there is likely to be comparable flexibility in heart growth or regeneration in the first weeks of human life.

Oxidative stress at birth leads to the production of chemically reactive molecules called reactive oxygen species (ROS) and induction of a DNA-damage response. Puente *et al.* identify these processes as the stimuli for cardiomyocyte cell-cycle withdrawal in a high-oxygen environment in newborn mice, and show that inhibition of oxidative stress through treatment with a ROS scavenger (*N*-acetylcysteine) prolongs the ability of the mouse heart to regenerate to 21 days after birth[4], compared with the normal 7-day window[3]. But even this window of proliferation closed after this point, indicating that inhibition of oxidative stress can delay, but not prevent, cardiomyocyte cell-cycle withdrawal. Delayed withdrawal for 1–2 weeks after birth has been observed in several mouse models, but in most cases the cells eventually stop dividing[2]. The overriding mechanism that ultimately prevents cardiomyocytes from proliferating is not fully understood and remains an area of intense research interest.

Although Puente and colleagues show that ROS scavenging can prolong the heart's regenerative capacity in newborn mice, they do not test whether this treatment promotes regeneration of the adult mouse heart after injury. Antioxidant therapy using the same ROS scavenger has been tested in humans as a means of reducing muscle damage after a heart attack[5], but limited benefit was observed, and this therapeutic approach is not in wide use. Because adult heart muscle is distinct from that of neonates in terms of metabolism and cell-cycle regulation, it seems unlikely that, under conditions of oxidative stress, adult cardiomyocytes will show an increased regenerative response to ROS scavenging. However, this is certainly an area worthy of future investigation.

It is interesting to speculate that the evolutionary transition to terrestrial life in atmospheric oxygen led to the inability of mammals to regenerate heart muscle. Zebrafish inhabit a relatively hypoxic environment and have a greater capacity to regenerate their hearts and limbs than mammals. Although it may not be possible to test the effects of atmospheric levels of oxygen or hyperoxia in fish, manipulation of oxygen levels in amphibians, which can also regenerate their hearts, could be used to test this hypothesis[3].

The definitive test might be to examine heart regeneration in terrestrial lizards, which can regenerate their tails (although the new tail is not exactly the same as the original)[6]. In mammals, it seems that exposure to atmospheric oxygen at birth leads to development of the more powerful muscle cells needed in a closed circulatory system, with the trade-off being a relative inability to regenerate heart muscle after injury or disease in adulthood. But further understanding of the transitions that occur in newborn mammals, such as that provided by Puente *et al.*, could reveal approaches by which to rejuvenate or repair diseased hearts. ■

Katherine E. Yutzey *is in the Division of Molecular Cardiovascular Biology, Cincinnati Children's Medical Center, Cincinnati, Ohio 45229, USA.*
e-mail: katherine.yutzey@cchmc.org

1. Xin, M., Olson, E. N. & Bassel-Duby, R. *Nature Rev. Mol. Cell Biol.* **14,** 529–541 (2013).
2. Kikuchi, K. & Poss, K. D. *Annu. Rev. Cell Dev. Biol.* **28,** 719–741 (2012).
3. Porrello, E. R. *et al. Science* **331,** 1078–1080 (2011).
4. Puente, B. N. *et al. Cell* **157,** 565–579 (2014).
5. Sochman, J. *J. Am. Coll. Cardiol.* **39,** 1422–1428 (2002).
6. Ritzman, T. B. *et al. Anat. Rec.* **295,** 1596–1608 (2012).

IMMUNOLOGY

To affinity and beyond

Tracking B cells in germinal centres — hotspots of B-cell proliferation and mutation during an immune response — reveals that those cells presenting the most antigen on their surface are programmed to dominate. SEE LETTER P.637

DAVID M. TARLINTON

On page 637 of this issue, Gitlin *et al.*[1] provide compelling insight into the processes driving the extraordinary increases that can occur in the affinity of antibodies for their target antigen molecules during vertebrate immune responses. The study resolves the question of how the rare B cells that produce high-affinity antibodies, which arise at random in a sea of average-affinity B cells, can dominate the response numerically within days. The answer is proliferation, but not through increased frequency or speed in the cells' basic response cycle of 'proliferate, mutate and select'. Instead, they are programmed by T cells, on the basis of the amount of antigen presented on the B-cell surface, to divide more times than lesser-affinity B cells in between rounds of affinity-based selection.

During an immune response, B cells undergo random mutations in their immunoglobulin variable-region (IgV) genes, which encode the antigen-binding region of antibody proteins. B cells in which this somatic (non-germline) mutation results in enhanced antibody affinity are selected for further rounds of proliferation[2]. This process of affinity maturation occurs in germinal centres (GCs) — regions of high cell density in secondary lymphoid organs, such as the lymph nodes and spleen, that form in the early stages of an immune response. The characterization of an enzyme known as activation-induced cytidine deaminase[3] provided a biochemical mechanism for the somatic mutation of IgV genes. Subsequently, experiments in many systems[4,5] suggested that affinity-based selection is probably driven by B-cell competition in the GC for interactions with follicular helper T cells (T_{FH} cells), which stimulate multiple signalling pathways[6], including those involved in survival, division, differentiation and migration. But how changes in the affinity of a B cell for antigen are translated into changes in relative representation of that cell in the immune response remained unclear.

Germinal centres form spatially distinct light and dark zones. B cells in the light zone are considered to be neither dividing nor undergoing IgV-gene mutation, but to be receiving signals from T_{FH} cells and thus being subjected to affinity-based selection. Dark-zone B cells, by contrast, are thought of as the survivors of previous rounds of selection, and to be actively proliferating and mutating. Affinity maturation occurs by B cells cycling

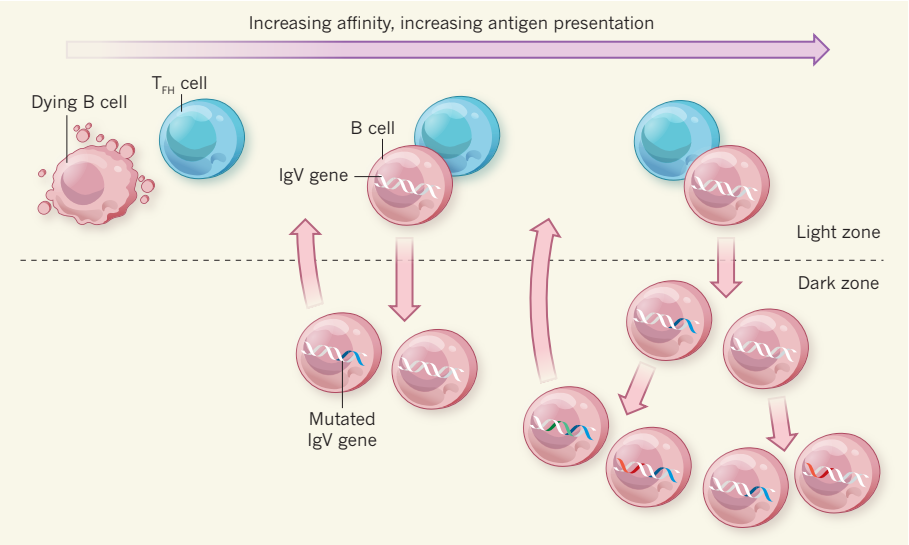


Figure 1 | Selection in the germinal centre. During an immune response to a specific antigen, B cells undergo antibody-affinity maturation in germinal centre (GC) regions of the lymphoid organs. This process involves the introduction of mutations in the B cells' IgV genes as the cells divide in the dark zone of the GC. The B cells then migrate to the GC light zone, where interactions with follicular helper T cells (T_{FH} cells) determine which cells re-enter the dark zone — B cells in which IgV-gene mutation leads to the generation of higher-affinity antibodies survive and migrate, whereas those with diminished affinity die. Gitlin *et al.*[1] show that the amount of antigen captured by a light-zone B cell and presented on its surface to T_{FH} cells determines the number of rounds of division and mutation that the cell subsequently undergoes in the dark zone, and the time taken for this, before re-migration. Because higher-affinity cells, which correlate with more mutations, capture more antigen, this results in a feed-forward mechanism, ensuring that these cells dominate the response.

from zone to zone, undergoing proliferation and mutation, and then selection, as they move between zones.

But what actually happens to a positively selected B cell in the light zone to promote its dominance? Is it subject to enhanced survival, increased movement between zones or enhanced proliferation that bypasses the cyclical routine? More pragmatically, how can affinity maturation be analysed, when it may occur at a low frequency with unpredictable timing, and when the high-affinity cells thus generated may be identifiable only retrospectively and as a population, making their behaviour around the time of selection difficult to follow?

Gitlin *et al.* address these very questions. First, they studied the effect, in mice, of delivering extra antigen to a small minority of B cells in established GCs. They observed increased proliferation of the boosted antigen-specific B cells, initially in the dark zone and to an extent proportional to the amount of antigen delivered. Although this clearly demonstrated that extra antigen enhances the proliferation of B cells in the dark zone, it did not distinguish between an acceleration of the normal cyclic migration of B cells between zones and an increased number of divisions per dark-zone B cell before migration to the light zone.

To address this issue, the authors isolated B cells from GCs and quantified the relative proportion in the 'actively dividing' or 'non-cycling' phases of the cell cycle. They found unboosted B cells in the actively dividing phases in both GC zones and observed that, once in the dark zone, the cells underwent, on average, two cell divisions. Boosting with antigen increased the proportion of actively dividing B cells in the dark zone from 60% to almost 90%, while halving the migration of cells from the dark zone to the light zone. Together, these findings show that B cells undergoing selection in the light zone can be programmed to divide in the dark zone a variable number of times on the basis of their previous interactions with T_{FH} cells; such interactions are, in turn, determined by the amount of antigen the B cell presents to the T_{FH} cells (Fig. 1). These *in vivo* observations are strikingly similar to the finding[7] that the number of divisions that B cells undergo *in vitro* is dictated by the strength of the signal passing through the CD40 pathway — one of the drivers of B-cell behaviour in GCs that is provided by T_{FH} cells.

To relate these observations to 'real-world' immune responses, Gitlin and colleagues applied their GC B-cell division tracking to cases in which antigen availability was not manipulated after the initial immunization. They found that the B cells undergoing the most proliferation had six times more affinity-enhancing mutations in their IgV genes than the least-proliferating cells. Cells undergoing the most proliferation also contained the highest number of somatic mutations. These findings confirmed the relationship between proliferation, mutation and affinity maturation.

The ultimate conclusion of this work is that affinity maturation works through a feed-forward mechanism, in which improved affinity begets a 'stronger' T_{FH}-cell signal that begets increased dark-zone B-cell proliferation and mutation, begetting further improvements in affinity, antigen acquisition and thus even more proliferation and mutation. Such a mechanism allows an ever more rapid expansion of high-affinity cells in the population, overwhelming both low-affinity and nonspecific B cells.

Is this the end of the road for GC biology? Thankfully, no. Although it provides extraordinary insight into the mechanics of B-cell dominance, Gitlin and colleagues' study leaves unexamined the means by which such favoured GC B cells extricate themselves from the GC and become circulating memory B cells and long-lived antibody-secreting cells in the bone marrow. There is evidence for stringent, affinity-based selection of GC B cells into the long-lived bone-marrow population[8], but how such selective interactions differ from those measured by the authors remains unclear.

Moreover, it has been shown[9] that interactions with T_{FH} cells mediate the differentiation of GC B cells into antibody-secreting cells, which cease affinity maturation and leave the GC to migrate to the bone marrow. So it is puzzling that when GC B cells are provided with almost unlimited amounts of T_{FH}-cell-derived signals, as in Gitlin and colleagues' study, they prefer to divide rather than differentiate. Perhaps this is an issue of kinetics, an incorrect hypothesis of what induces differentiation or an indication of nuanced types of signalling through T_{FH} cells or other, regulatory, T cells. Resolving such issues will further our understanding of how the orchestration of B-cell behaviour in GCs helps to optimize immediate and long-term immune protection. ■

David M. Tarlinton *is at the Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia, and in the Department of Medical Biology, University of Melbourne, Parkville.*
e-mail: tarlinton@wehi.edu.au

1. Gitlin, A. D., Shulman, Z. & Nussenzweig, M. C. *Nature* **509,** 637–640 (2014).
2. MacLennan, I. C. *Annu. Rev. Immunol.* **12,** 117–139 (1994).
3. Muramatsu, M. *et al. Cell* **102,** 553–563 (2000).
4. Victora, G. D. & Nussenzweig, M. C. *Annu. Rev. Immunol.* **30,** 429–457 (2012).
5. Allen, C. D. C. *et al. Nature Immunol.* **5,** 943–952 (2004).
6. Crotty, S. *Annu. Rev. Immunol.* **29,** 621–663 (2011).
7. Turner, M. L., Hawkins, E. D. & Hodgkin, P. D. *J. Immunol.* **181,** 374–382 (2008).
8. Smith, K. G. C., Light, A., Nossal, G. J. V. & Tarlinton, D. M. *EMBO J.* **16,** 2996–3006 (1997).
9. Tarlinton, D. & Good-Jacobson, K. *Science* **341,** 1205–1211 (2013).

CORRECTION

In the News & Views article 'Sensory biology: Radio waves zap the biomagnetic compass' by Joseph L. Kirschvink (*Nature* **509,** 296–297; 2014), Figure 1a was wrongly credited. It should have been credited to Marianne Hanzlik.

ARTICLE

doi:10.1038/nature13302

A draft map of the human proteome

Min-Sik Kim[1,2], Sneha M. Pinto[3], Derese Getnet[1,4], Raja Sekhar Nirujogi[3], Srikanth S. Manda[3], Raghothama Chaerkady[1,2], Anil K. Madugundu[3], Dhanashree S. Kelkar[3], Ruth Isserlin[5], Shobhit Jain[5], Joji K. Thomas[3], Babylakshmi Muthusamy[3], Pamela Leal-Rojas[1,6], Praveen Kumar[3], Nandini A. Sahasrabuddhe[3], Lavanya Balakrishnan[3], Jayshree Advani[3], Bijesh George[3], Santosh Renuse[3], Lakshmi Dhevi N. Selvan[3], Arun H. Patil[3], Vishalakshi Nanjappa[3], Aneesha Radhakrishnan[3], Samarjeet Prasad[1], Tejaswini Subbannayya[3], Rajesh Raju[3], Manish Kumar[3], Sreelakshmi K. Sreenivasamurthy[3], Arivusudar Marimuthu[3], Gajanan J. Sathe[3], Sandip Chavan[3], Keshava K. Datta[3], Yashwanth Subbannayya[3], Apeksha Sahu[3], Soujanya D. Yelamanchi[3], Savita Jayaram[3], Pavithra Rajagopalan[3], Jyoti Sharma[3], Krishna R. Murthy[3], Nazia Syed[3], Renu Goel[3], Aafaque A. Khan[3], Sartaj Ahmad[3], Gourav Dey[3], Keshav Mudgal[7], Aditi Chatterjee[3], Tai-Chung Huang[1], Jun Zhong[1], Xinyan Wu[1,2], Patrick G. Shaw[1], Donald Freed[1], Muhammad S. Zahari[2], Kanchan K. Mukherjee[8], Subramanian Shankar[9], Anita Mahadevan[10,11], Henry Lam[12], Christopher J. Mitchell[1], Susarla Krishna Shankar[10,11], Parthasarathy Satishchandra[13], John T. Schroeder[14], Ravi Sirdeshmukh[3], Anirban Maitra[15,16], Steven D. Leach[1,17], Charles G. Drake[16,18], Marc K. Halushka[15], T. S. Keshava Prasad[3], Ralph H. Hruban[15,16], Candace L. Kerr[19]†, Gary D. Bader[5], Christine A. Iacobuzio-Donahue[15,16,17], Harsha Gowda[3] & Akhilesh Pandey[1,2,3,4,15,16,20]

The availability of human genome sequence has transformed biomedical research over the past decade. However, an equivalent map for the human proteome with direct measurements of proteins and peptides does not exist yet. Here we present a draft map of the human proteome using high-resolution Fourier-transform mass spectrometry. In-depth proteomic profiling of 30 histologically normal human samples, including 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells, resulted in identification of proteins encoded by 17,294 genes accounting for approximately 84% of the total annotated protein-coding genes in humans. A unique and comprehensive strategy for proteogenomic analysis enabled us to discover a number of novel protein-coding regions, which includes translated pseudogenes, non-coding RNAs and upstream open reading frames. This large human proteome catalogue (available as an interactive web-based resource at http://www.humanproteomemap.org) will complement available human genome and transcriptome data to accelerate biomedical research in health and disease.

Analysis of the complete human genome sequence has thus far led to the identification of approximately 20,687 protein-coding genes[1], although the annotation still continues to be refined. Mass spectrometry has revolutionized proteomics studies in a manner analogous to the impact of next-generation sequencing on genomics and transcriptomics[2–4]. Several groups, including ours, have used mass spectrometry to catalogue complete proteomes of unicellular organisms[5–7] and to explore proteomes of higher organisms, including mouse[8] and human[9,10]. To develop a draft map of the human proteome by systematically identifying and annotating protein-coding genes in the human genome, we carried out proteomic profiling of 30 histologically normal human tissues and primary cells using high-resolution mass spectrometry. We generated tandem mass spectra corresponding to proteins encoded by 17,294 genes, accounting for approximately 84% of the annotated protein-coding genes in the human genome—to our knowledge the largest coverage of the human proteome reported so far. This includes mass spectrometric evidence for proteins encoded by 2,535 genes that have not been previously observed as evidenced by their absence in large community-based proteomic data sets—PeptideAtlas[11], GPMDB[12] and neXtProt[13] (which includes annotations from the Human Protein Atlas[14]).

A general limitation of current proteomics methods is their dependence on predefined protein sequence databases for identifying proteins. To overcome this, we also used a comprehensive proteogenomic analysis strategy to identify novel peptides/proteins that are currently not part of annotated protein databases. This approach revealed novel protein-coding genes in the human genome that are missing from current genome annotations in addition to evidence of translation of several annotated pseudogenes as well as non-coding RNAs. As discussed below, we provide evidence for revising hundreds of entries in protein databases based on our data. This includes novel translation start sites, gene/exon extensions and novel coding exons for annotated genes in the human genome.

Generating a high-quality mass spectrometry data set

To generate a baseline proteomic profile in humans, we studied 30 histologically normal human cell and tissue types, including 17 adult tissues,

[1]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [2]Department of Biological Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [3]Institute of Bioinformatics, International Tech Park, Bangalore 560066, India. [4]Adrienne Helis Malvin Medical Research Foundation, New Orleans, Louisiana 70130, USA. [5]The Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada. [6]Department of Pathology, Universidad de La Frontera, Center of Genetic and Immunological Studies-Scientific and Technological Bioresource Nucleus, Temuco 4811230, Chile. [7]School of Medicine, Imperial College London, South Kensington Campus, London SW7 2AZ, UK. [8]Department of Neurosurgery, Postgraduate Institute of Medical Education & Research, Chandigarh 160012, India. [9]Department of Internal Medicine Armed Forces Medical College, Pune 411040, India. [10]Department of Neuropathology, National Institute of Mental Health and Neurosciences, Bangalore 560029, India. [11]Human Brain Tissue Repository, Neurobiology Research Centre, National Institute of Mental Health and Neurosciences, Bangalore 560029, India. [12]Department of Chemical and Biomolecular Engineering and Division of Biomedical Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. [13]Department of Neurology, National Institute of Mental Health and Neurosciences, Bangalore 560029, India. [14]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21224, USA. [15]The Sol Goldman Pancreatic Cancer Research Center, Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA. [16]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA. [17]Department of Surgery, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA. [18]Departments of Immunology and Urology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA. [19]Department of Obstetrics and Gynecology, Johns Hopkins University School of Medicine Baltimore, Maryland 21205, USA. [20]Diana Helis Henry Medical Research Foundation, New Orleans, Louisiana 70130, USA. †Present address: Department of Biochemistry and Molecular Biology, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA.

7 fetal tissues, and 6 haematopoietic cell types (Fig. 1a). Pooled samples from three individuals per tissue type were processed and fractionated at the protein level by SDS–polyacrylamide gel electrophoresis (SDS–PAGE) and at the peptide level by basic reversed-phase liquid chromatography (RPLC) and analysed on high-resolution Fourier-transform mass spectrometers (LTQ-Orbitrap Elite and LTQ-Orbitrap Velos) (Fig. 1b). To generate a high-quality data set, both precursor ions and higher-energy collisional dissociation (HCD)-derived fragment ions were measured using the high-resolution and high-accuracy Orbitrap mass spectrometer. Approximately 25 million high-resolution tandem mass spectra, acquired from more than 2,000 LC-MS/MS (liquid chromatography followed by tandem mass spectrometry) runs, were searched against NCBI's RefSeq[15] human protein sequence database using the MASCOT[16] and SEQUEST[17] search engines. The search results were rescored using the Percolator[18] algorithm and a total of approximately 293,000 non-redundant peptides were identified at a q value < 0.01 with a median mass measurement error of approximately 260 parts per billion (Extended Data Fig. 1a). The median number of peptides and corresponding tandem mass spectra identified per gene are 10 and 37, respectively, whereas the median protein sequence coverage was approximately 28% (Extended Data Fig. 1b, c). It should be noted, however, that false-positive rates for subgroups of peptide-spectrum matches can vary upon the nature of peptides, such as their size, the charge state of precursor peptide ions or missed enzymatic cleavage (Extended Data Fig. 1d–f and Supplementary Information).

We compared our data set with two of the largest human peptide-based resources, PeptideAtlas and GPMDB. These two databases contain curated peptide information that has been collected from the proteomics community over the past decade. Notably, almost half of the peptides we identified were not deposited in either one of these resources. Also, the novel peptides in our data set constitute 37% of the peptides in PeptideAtlas and 54% of peptides in the case of GPMDB (Extended Data Fig. 1g, h). This marked increase in the coverage of human proteomic data was made possible by the breadth and depth of our analysis as most of the cells and tissues that we analysed have not previously been studied using similar methods. The depth of our analysis enabled us to identify protein products derived from two-thirds (2,535 out of 3,844) of proteins designated as 'missing proteins'[19] for lack of protein-based evidence. Several hypothetical proteins that we identified have a broad tissue distribution, indicating the inadequate sampling of the human proteome thus far (Extended Data Fig. 2a).

Landscape of protein expression in cells and tissues

On the basis of gene expression studies, it is clear that there are several genes that are involved in basic cellular functions that are constitutively expressed in almost all the cells/tissues. Although the concept of 'housekeeping genes' as genes that are expressed in all tissues and cell types is widespread among biologists, there is no readily available catalogue of such genes. Moreover, the extent to which these transcripts are translated into proteins remains unknown. We detected proteins encoded by 2,350 genes across all human cells/tissues with these highly abundant 'housekeeping proteins' constituting approximately 75% of total protein mass based on spectral counts (Extended Data Fig. 2b). The large majority of these highly expressed housekeeping proteins include histones, ribosomal proteins, metabolic enzymes and cytoskeletal proteins. One of the caveats of tissue proteomics is the contribution of vasculature, blood and haematopoietic cells. Thus, proteins designated as housekeeping proteins based on analysis of tissue proteomes could be broadly grouped into two categories, those that are truly expressed in every single cell type and those that are found in every tissue (for example, endothelial cells). Another caveat to be noted here is that some proteins that are indeed expressed in all tissues might not be detected in some of the tissues because of inadequate sampling by mass spectrometry. Thus, this list of housekeeping proteins will continue to be refined as additional in-depth analyses are carried out.

We used a label-free method based on spectral counting to quantify protein expression across cells/tissues. Although more variable as compared

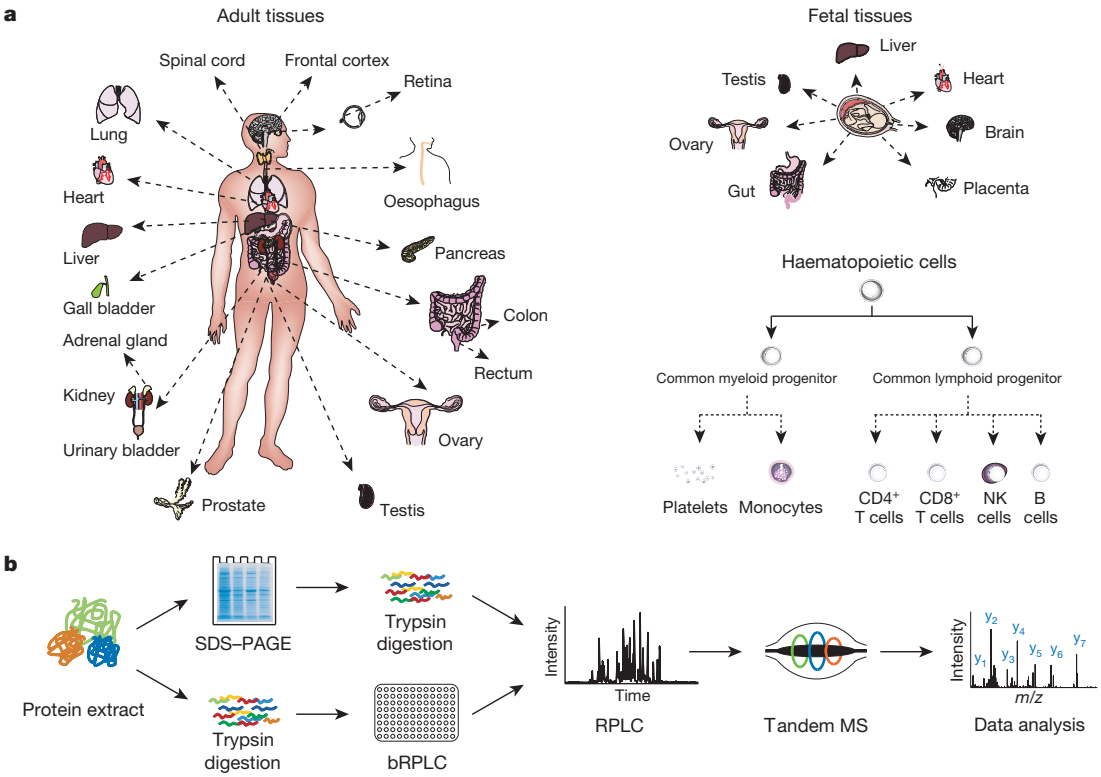


Figure 1 | Overview of the workflow and comparison of data with public repositories. a, The adult/fetal tissues and haematopoietic cell types that were analysed to generate a draft map of the normal human proteome are shown. **b**, The samples were fractionated, digested and analysed on the high-resolution and high-accuracy Orbitrap mass analyser as shown. Tandem mass spectrometry data were searched against a known protein database using SEQUEST and MASCOT database search algorithms.

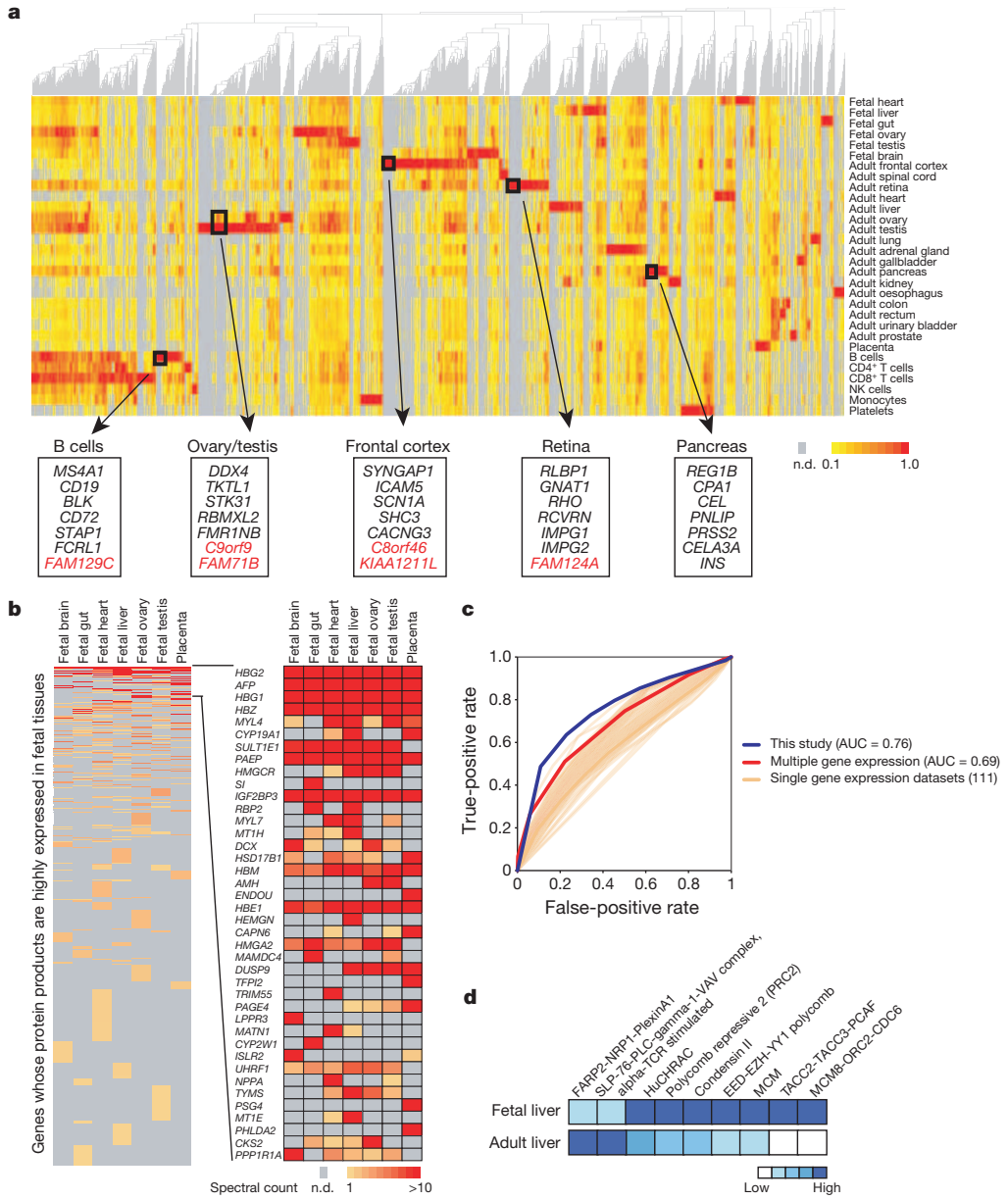


Figure 2 | Landscape of the normal human proteome. **a**, Tissue-supervised hierarchical clustering reveals the landscape of gene expression across the analysed cells and tissues. Selected tissue-restricted genes are highlighted in boxes to show some well-studied genes (black) as well as hypothetical proteins of unknown function (red). The colour key indicates the normalized spectral counts per gene detected across the tissues. n.d., not determined. **b**, A heat map showing tissue expression of fetal tissue-restricted genes ordered by average expression across fetal tissues (left) and a zoom-in of the top 40 most abundant genes (right). The colour key indicates the spectral counts per gene. **c**, An ROC curve showing a comparison of the performance of the current data set (blue, area under the curve (AUC) = 0.762) with 111 individual gene expression data sets (orange) and a composite of the 111 individual data sets (red, AUC = 0.692). **d**, Developmental stage-specific differential expression of protein complexes in fetal and adult liver tissues. Heat map shows protein complexes with less than or equal to half of their subunits expressed in one of the tissue types. The darker the colour, the greater the number of expressed subunits.

to label-based methods, this method is readily applicable to analysis of a large number of samples[8] and has been shown to be reproducible[20]. Supervised hierarchical clustering showed proteins encoded by some genes to be expressed in only a few cells/tissues, whereas others were more broadly expressed (Fig. 2a). Some proteins detected in only one sample were encoded by well-known genes like *CD19* in B cells, *SCN1A* in the frontal cortex and *GNAT1* in the retina, whereas others were encoded by ill-characterized genes. For example, *C8orf46* was expressed in the adult frontal cortex whereas *C9orf9* was expressed in adult ovary and testis. Overall, we detected proteins encoded by 1,537 genes only in one of 30 human samples examined in this study (Extended Data Fig. 2c). These may or may not be tissue-specific genes because of the limit of detection of mass spectrometry and because this analysis did not sample every human cell or tissue type. Because methods based on antibody-based detection can be more sensitive, we performed western blotting experiments to confirm the tissue-restricted expression pattern of some of the proteins against which appropriate antibodies were available. Of 32 proteins tested, eight proteins exhibited a tissue-specific expression in agreement with mass spectrometry-derived data (Extended Data Fig. 3a). Four proteins exhibited a more widespread expression, although in each of these cases extra bands were detected (Extended Data Fig. 3b). In eighteen cases, the antibody did not recognize a protein in the expected size range at all, and no band was detectable in the remaining two cases.

A number of proteins are expressed during development in fetal tissues but not in normal adult tissues. Although earlier studies have focused on a few fetal tissues like fetal brain[21] or liver[22], our study provides the first general survey of the fetal proteome. We detected proteins encoded by 735 genes that are expressed more than tenfold in fetal samples compared to adult tissues/cells. A heat map highlights the expression level of putative fetal-tissue-restricted genes across various fetal tissues (Fig. 2b). The list includes the well-known oncofetal antigens, alpha fetoprotein (*AFP*) and insulin-like growth factor-2 binding protein-3/IGF-2 mRNA binding protein-3 (*IGF2BP3*). High levels of AFP in serum and cerebrospinal fluid are clinically used as biomarkers for neural tube defects, teratoma and yolk sac tumours. Some of the proteins expressed during development in ovary and testis can serve as potential biological markers for identifying cancers of different lineages in the future.

In the past, gene expression profiles across various experimental conditions or tissues have been used to investigate the likelihood of co-expressed genes to physically interact at the protein level[23,24]. We proposed

that protein expression patterns should be a better predictor of protein–protein interactions than gene expression measured at the messenger RNA level. We correlated the expression level for each available protein pair across all 30 cells/tissues in our data using Pearson correlation and compared this to known protein–protein interactions. We then repeated this analysis using correlations obtained from 111 published gene expression data sets. Receiver operating characteristic (ROC) curve analysis clearly shows that data from the human proteome map outperforms that from gene expression profiles for predicting protein–protein interactions, even if all the gene expression data sets are combined and used as a single predictor (Fig. 2c). It should be pointed out, however, that although the use of protein expression data are useful for predicting protein–protein interactions, it is unlikely by itself to be sufficient for such predictions.

Many proteins interact with different partners in different tissues or at different stages of development, although they are not traditionally studied in this fashion. To investigate tissue-restricted expression of protein complex subunits, we evaluated 938 complexes with three or more subunits from the CORUM database[25] and found 679 protein complex subunits showing differential expression in at least one of the 30 tissues. In contrast, there were 34 complexes where all the subunits were expressed concurrently in all tissues. Interestingly, there were 201 instances where differential expression of subunits of complexes was observed across the adult and fetal tissues, indicating that these complexes are dynamic and probably have distinct composition during ontogeny. This dynamic composition is probably related to developmental stage-specific processes in which these complexes are involved. For example, mini-chromosome maintenance (MCM) complex components are highly expressed in fetal liver in contrast to FARP2–NRP1–PLXNA1 complex members, which are highly expressed in the adult liver (Fig. 2d).

Detecting protein isoforms

Alternative splicing gives rise to a large number of splice variants at the RNA level, some of which can encode distinct protein isoforms. Multiple protein isoforms are contributed by only one-third of annotated genes, whereas the remaining two-thirds generate only a single protein product according to the RefSeq database[15] (Extended Data Fig. 2d). Although our primary goal was not to obtain complete coverage of splice isoforms, we identified isoform-specific peptides for 2,861 protein isoforms derived from 2,450 genes. For example, we detected isoform 1 of FYN protein tyrosine kinase in brain and isoform 2 in haematopoietic cells (Fig. 3a). This is significant because although we did not detect the third known isoform of *FYN*, the two isoforms that we identified are known to possess distinct functional properties.

An interactive portal for exploring the human proteome

We have developed a portal for the Human Proteome Map (http://www.humanproteomemap.org) that makes it possible to test and generate hypotheses regarding gene families, protein complexes, signalling pathways, biomarkers, therapeutic targets, immune function and human development. As an illustration, one can explore the protein components of the 20S constitutive proteasome and immunoproteasome complexes (Fig. 3b and Extended Data Fig. 2e). Three of the subunits in 20S constitutive proteasome (PSMB5, PSMB6 and PSMB7; coloured red in Fig. 3b left panel) are known to be replaced by three other subunits (PSMB8, PSMB9 and PSMB10; coloured green) in the 20S immunoproteasome[26]. As shown in the heat map from Human Proteome Map, PSMB8, PSMB9 and PSMB10 are highly expressed in immune cell types.

Novel protein-coding regions in the human genome

The evidence for protein-coding potential is still largely driven by gene prediction algorithms or complementary DNAs and does not routinely include direct detection/measurement of proteins/peptides. Following protein database search of our large-scale proteome analysis, we extracted approximately 16 million MS/MS spectra that did not match currently annotated proteins. These unmatched spectra were searched using a

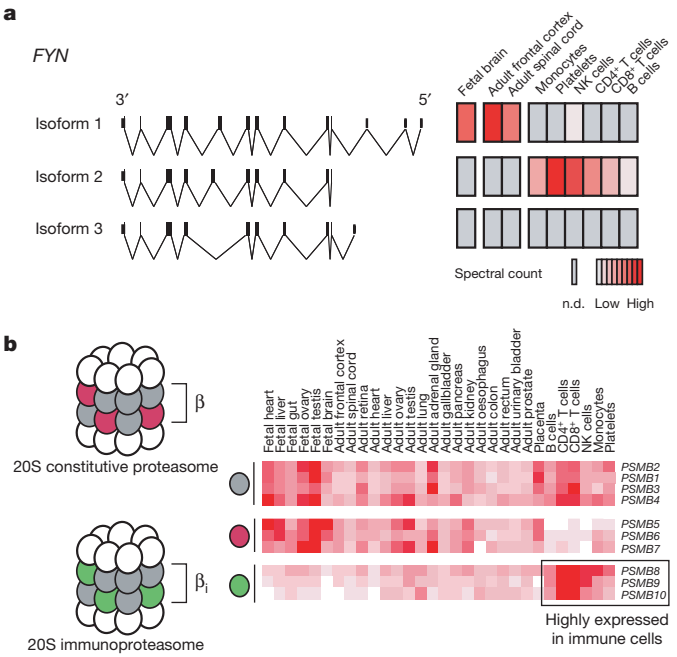


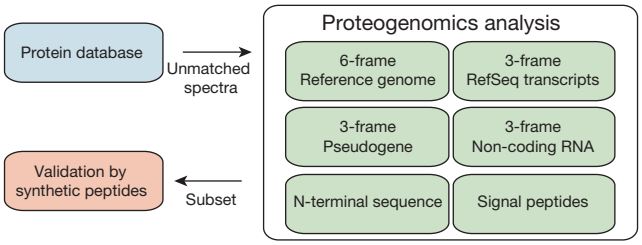
Figure 3 | Isoform-specific expression. **a**, Exon structure of three known isoforms of *FYN* (left) along with abundance of isoform-specific peptides detected in the indicated cells/tissues (right). The colour key indicates a relative expression based on the spectral counts of isoform-specific peptides detected. **b**, 20S constitutive proteasome and 20S immunoproteasome core complexes. Expression of their corresponding components are depicted by a heat map (red indicates higher expression) in the Human Proteome Map portal.

unique proteogenomic analysis strategy developed by our group against conceptually translated human reference genome, RefSeq transcript sequences, non-coding RNAs and pseudogenes. In addition, the data were searched against theoretical protein amino termini and predicted signal sequences (Fig. 4a). As a result, we confirmed translation of 17,294 annotated protein-coding genes that include 4,105 protein N termini, 223,385 exons and 66,947 splice junctions (Fig. 4b). More importantly, we discovered 808 novel annotations of the human genome including translation of 140 pseudogenes, 44 novel ORFs, 106 novel coding regions/exons within annotated gene structures and 110 gene/protein/exon extension events in addition to 198 novel protein N termini and 201 novel signal peptide cleavage sites (Fig. 4b and Supplementary Table 1). Although the peptides were identified at less than 1% false discovery rate (FDR) with parts per billion median mass measurement error (Extended Data Fig. 1i), we carried out an additional level of manual inspection of the MS/MS spectra to reduce false positives in this type of analysis[27]. In addition, because these novel findings are based on matches of tandem mass spectra to translated genomic/transcriptomic sequences, we experimentally confirmed a subset of the identifications from various proteogenomic categories through matching of MS/MS spectra from 98 synthetic peptides with those obtained from analysis of human cells/tissues (Supplementary Data).

Upstream and other ORFs

We identified 28 genes with upstream ORFs (uORFs) where peptides mapped to 5′ untranslated regions (UTRs) of known human RefSeq transcripts. For example, we identified peptides that mapped to the 5′ UTR of solute carrier family 35, member A4 transcript (*SLC35A4*). Although we did not identify any unique peptides from the protein encoded by *SLC35A4*, which is probably related to its low abundance or cell/tissue-specific expression, we identified four peptides derived from the protein encoded by the uORF in 25 out of 30 samples (Extended Data Fig. 5a). This uORF encodes a 103-amino-acid protein that contains a transmembrane region within a putative BNIP3 domain (BCL2/adenovirus

a



b

	Number of genes involved or annotation confirmed /added/altered
Genes whose products were detected	17,294
Confirmed exons	223,385
Confirmed N termini	4,105
Confirmed exon–exon junctions	66,947
Signal peptide cleavage site	329
Confirmation of annotated sites	128
Unannotated cleavage sites	201
Novel protein-coding regions	193
Pseudogenes	140
Non-coding RNAs	9
Upstream ORFs	29
Other ORFs	15
Novel coding regions/exons	106
Novel N termini	198
Gene/protein extensions	70
N-terminal extension	58
C-terminal extension	12
Exon extension	40

Figure 4 | Proteogenomic analysis. **a**, An overview of the multiple databases used in the proteogenomic analysis. A subset of peptides corresponding to genome search-specific peptides were synthesized and analysed by mass spectrometry. **b**, Overall summary of the results from the current study.

E1B 19 kDa protein-interacting protein 3) that is required for mitochondrial localization, suggesting that this protein may play a role in mitochondrial function.

We identified eight cases where we observed peptides that mapped to an ORF located in an alternate reading frame within coding regions of annotated genes. For example, we identified peptides that mapped to a novel ORF of 159 amino acids within the *C11orf48* gene. The protein encoded by the *C11orf48* gene was identified only in the adult retina, although we identified three peptides encoded by the novel ORF from 17 various cells/tissues. We also identified peptide matches to seven ORFs located within 3′UTRs. As an example, a novel ORF comprising of 524 amino acids in the 3′UTR of the *CHTF8* gene was identified on the basis of multiple peptides. The translation initiation site of this novel ORF overlaps the stop codon of the *CHTF8* gene (Extended Data Fig. 4a). Remarkably, the protein encoded by this novel ORF was observed in haematopoietic cells where we did not detect the CHTF8 protein. In addition, this novel gene product was expressed at higher levels in fetal ovary and adult testis than the protein encoded by *CHTF8*. These observations suggest that the translational control for these two proteins encoded by the same gene structure is likely to be different. We also identified a peptide encoded by an ORF within a human endogenous retrovirus (Extended Data Fig. 5b). Domain analysis revealed the presence of a signal peptide at the N terminus along with other domains including Furin-like repeats. In fact, during preparation of this manuscript, a report was published in which this protein was designated as suppressyn and shown to inhibit cell–cell fusion in trophoblast cells[28].

Non-coding RNAs and pseudogenes

We detected several peptides that were not mapped to RefSeq proteins but corresponded to nine transcripts annotated as non-coding RNAs. As shown in Extended Data Fig. 4b, five peptides (including one across an exon-exon junction) were mapped to a region on chromosome 1 that was annotated as a non-coding RNA. Although long non-coding RNAs are, by definition, not supposed to encode proteins[29], our data indicate that some of these encode proteins, which can be discovered by deep proteomic profiling of cells/tissues.

A few studies have recently examined the transcription of pseudogenes on a global scale and found approximately 2,000 pseudogenes to be transcribed either ubiquitously or in a tissue/cancer-specific manner[30,31]. In our data set, we identified more than 200 peptides that are encoded by 140 pseudogenes. These 140 pseudogenes originate from 110 parental genes. To derive this unambiguous set, we filtered out several peptides for which sequence change could be explained by SNPs reported for corresponding genes in dbSNP. When we looked at the tissue distribution of the translated pseudogenes, we observed that roughly half of the pseudogenes were translated in a cell/tissue-restricted manner whereas a small minority was expressed globally (Fig. 5a), a pattern similar to that described for pseudogene transcripts[30,31]. For instance, VDAC1P7 was found to be translated globally (22 of 30 cells/tissues analysed) whereas MAGEB6P1 was detected only in adult testes although its parental gene, *EIF4B*, is widely expressed (Fig. 5a). Extended Data Fig. 5c shows two peptides that map to *MAGEB6P1*, a pseudogene, along with an alignment of the identified peptides with the corresponding peptides in the parental gene. It should be noted that there is still a small chance that some SNPs are not represented in our analysis because we did not analyse the genomes of these individuals. However, this is unlikely to affect our results greatly because it has been estimated that approximately 98% SNPs with 1% frequency have been detected from the 1000 Genomes Project[32].

Protein N termini and signal peptides

We confirmed 4,105 annotated N termini in RefSeq (from 4,132 protein-coding genes) to generate the largest catalogue of experimentally validated translational start sites in humans (~20% of annotated genes). Annotation of start sites in proteins can be imprecise because initiation AUG codons are generally predicted by computation and are prone to errors. For instance, the optimal Kozak consensus sequence (CCACC [AUG]G), which is widely believed to surround the true initiation codon, occurs only in half of the human genes[33]. By searching unmatched MS/MS spectra against customized databases, we identified 3 cases that map upstream and 195 cases that map downstream (within 100 amino acids) of the currently annotated translational initiation sites (Fig. 5b). The nucleotide sequences that surround the AUG codon of these novel translational start sites show the same pattern as bona fide translational start sites (Extended Data Fig. 6), strongly indicating that we have identified true novel or alternative start sites. We also confirmed 201 annotated signal peptide cleavage sites and revised the predictions in 128 other cases (Supplementary Information).

Discussion

Here we present a high-coverage map (covering >84% of human protein-coding genes) of the human proteome using high-resolution mass spectrometry. This demonstrates the feasibility of comprehensively exploring the human proteome on an even larger scale in the near future such as the initiative recently announced by the Human Proteome Organization; the Chromosome-Centric Human Proteome Project[34–36]. Our findings also highlight the need for using direct protein sequencing technologies like mass spectrometry to complement genome annotation efforts. This process can be facilitated by workflows suitable for proteogenomic analyses. Even greater depth can lead to an increase in sequence coverage by employing methodologies such as multiple proteases, direct capture of N termini, enrichment of post-translationally modified peptides, exhaustive fractionation and incorporating alternative sequencing technologies such as electron transfer dissociation (ETD) and top-down mass spectrometry. Finally, these strategies can be further complemented by sampling of individual cell types of human tissues/organs to develop a 'human cell map'. With the availability of both genomic and proteomic landscapes, integrating the information from both resources is likely to accelerate

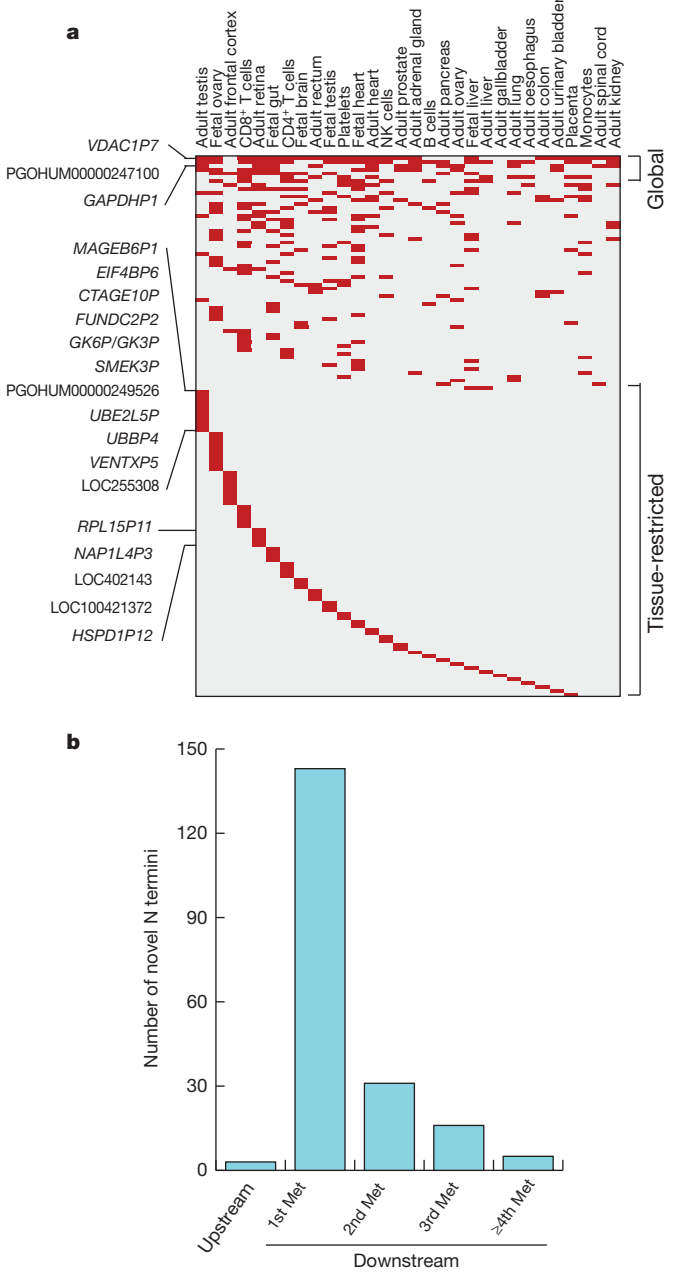


Figure 5 | Translation of pseudogenes and identification of novel N termini. **a**, A heat map shows the expression of pseudogenes (listed by name or accession code number) across the analysed cells/tissues. Some pseudogenes such as *VDAC1P7* and *GAPDHP1* were found to be globally expressed, whereas others were more restricted in their expression or were detected only in a single cell type/tissue as indicated. **b**, The distribution of novel N termini detected with N-terminal acetylation is shown with respect to the location of the annotated translational start site. All sites in the 5′ UTR are labelled upstream whereas those located downstream of the annotated AUG start sites are labelled as 1st Met, 2nd Met and so on.

basic as well as translational research in the years to come through a better understanding of gene-protein-pathway networks in health and disease.

METHODS SUMMARY

Protein was extracted from sorted cells or tissue/organs from three individuals and pooled before processing. The complex mixture of peptides obtained from tryptic digestion of gel bands from SDS–PAGE or basic RPLC fractions was separated on a micro-capillary RPLC column, ionized by nanoflow electrospray ionization and analysed on LTQ-Orbitrap Elite or LTQ-Orbitrap Velos mass spectrometers using HCD fragmentation. Raw mass spectrometry files were processed in Proteome Discoverer and peak lists were first searched against a human RefSeq protein sequence database. The unmatched spectra were subsequently searched against a series of custom polypeptide sequence databases. Only those genome-search-specific peptides that had unique hits in the human genome were investigated further for proteogenomic analysis. A subset of genome-search-specific peptides was generated as synthetic peptides and their fragmentation pattern was compared with corresponding tandem mass spectra identified from proteogenomic analysis. Hierarchical clustering was carried out using normalized spectral counting-based quantification values. Full methods and associated references are available in the online version of this paper.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 August 2013; accepted 31 March 2014.

1. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
2. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
3. Bensimon, A., Heck, A. J. & Aebersold, R. Mass spectrometry-based proteomics and network biology. *Annu. Rev. Biochem.* **81,** 379–405 (2012).
4. Cravatt, B. F., Simon, G. M. & Yates, J. R. III. The biological impact of mass-spectrometry-based proteomics. *Nature* **450,** 991–1000 (2007).
5. Nagaraj, N. *et al.* System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11,** M111.013722 (2012).
6. Picotti, P. *et al.* A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **494,** 266–270 (2013).
7. Kelkar, D. S. *et al.* Proteogenomic analysis of *Mycobacterium tuberculosis* by high resolution mass spectrometry. *Mol. Cell. Proteomics* **10,** M111.011627 (2011).
8. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143,** 1174–1189 (2010).
9. Gholami, A. M. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell Rep.* **4,** 609–620 (2013).
10. Branca, R. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature Methods* **11,** 59–62 (2014).
11. Farrah, T. *et al.* The state of the human proteome in 2012 as viewed through PeptideAtlas. *J. Proteome Res.* **12,** 162–171 (2013).
12. Craig, R., Cortens, J. P. & Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3,** 1234–1242 (2004).
13. Gaudet, P. *et al.* neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **12,** 293–298 (2013).
14. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nature Biotechnol.* **28,** 1248–1250 (2010).
15. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42,** D756–D763 (2014).
16. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).
17. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).
18. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4,** 923–925 (2007).
19. Lane, L. *et al.* Metrics for the human proteome project 2013–2014 and strategies for finding missing proteins. *J. Proteome Res.* **13,** 15–20 (2014).
20. Mosley, A. L. *et al.* Highly reproducible label free quantitative proteomic analysis of RNA polymerase complexes. *Mol. Cell. Proteomics* **10,** M110.000687 (2011).
21. Fountoulakis, M., Juranville, J. F., Dierssen, M. & Lubec, G. Proteomic analysis of the fetal brain. *Proteomics* **2,** 1547–1576 (2002).
22. Ying, W. *et al.* A dataset of human fetal liver proteome identified by subcellular fractionation and multiple protein separation and identification technology. *Mol. Cell. Proteomics* **5,** 1703–1707 (2006).
23. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12,** 37–46 (2002).
24. Ge, H., Liu, Z., Church, G. M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.* **29,** 482–486 (2001).
25. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.* **38,** D497–D501 (2010).
26. Ferrington, D. A. & Gregerson, D. S. Immunoproteasomes: structure, function, and antigen presentation. *Prog. Mol. Biol. Transl. Sci.* **109,** 75–112 (2012).
27. Steen, H. & Mann, M. The abc's (and xyz's) of peptide sequencing. *Nature Rev. Mol. Cell Biol.* **5,** 699–711 (2004).
28. Sugimoto, J., Sugimoto, M., Bernstein, H., Jinno, Y. & Schust, D. A novel human endogenous retroviral protein inhibits cell-cell fusion. *Sci. Rep.* **3,** 1462 (2013).
29. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154,** 240–251 (2013).
30. Kalyana-Sundaram, S. *et al.* Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **149,** 1622–1634 (2012).
31. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13,** R51 (2012).

32. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
33. Peri, S. & Pandey, A. A reassessment of the translation initiation codon in vertebrates. *Trends Genet.* **17,** 685–687 (2001).
34. Legrain, P. *et al.* The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **10,** M111.009993 (2011).
35. Paik, Y. K. *et al.* The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nature Biotechnol.* **30,** 221–223 (2012).
36. Marko-Varga, G., Omenn, G. S., Paik, Y. K. & Hancock, W. S. A first step toward completion of a genome-wide characterization of the human proteome. *J. Proteome Res.* **12,** 1–5 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to acknowledge the National Development and Research Institutes for some of the tissues. We acknowledge the assistance of V. Sandhya, V. Puttamallesh, U. Guha and B. Cole for help with analysis of some of the samples. We thank L. Lane and B. Amos for their assistance with the list of missing genes. This work was supported by an NIH roadmap grant for Technology Centers of Networks and Pathways (U54GM103520), NCI's Clinical Proteomic Tumor Analysis Consortium initiative (U24CA160036), a contract (HHSN268201000032C) from the National Heart, Lung and Blood Institute and the Sol Goldman Pancreatic Cancer Research Center. The authors acknowledge the joint participation by the Adrienne Helis Malvin Medical Research Foundation and the Diana Helis Henry Medical Research Foundation through its direct engagement in the continuous active conduct of medical research in conjunction with The Johns Hopkins Hospital and the Johns Hopkins University School of Medicine and the Foundation's Parkinson's Disease Programs. The analysis work was partially supported by the National Resource for Network Biology (P41GM103504). A.Mah., S.K.Sh., P.S. and T.S.K.P. are supported by DBT Program Support on Neuroproteomics (BT/01/COE/08/05) to IOB and NIMHANS. H.G. is a Wellcome Trust-DBT India Alliance Early Career Fellow. We thank Council of Scientific and Industrial Research, University Grants Commission and Department of Science and Technology, Government of India for research fellowships for S.M.P., R.S.N., A.R., M.K., G.J.S., S.C., P.R., J.S., S.S.M., D.S.K., S.R., S.K.Sr., K.K.D., Y.S., A.S., S.D.Y., N.S., S.A. and G.D.

Author Contributions A.P., H.G., R.C., M.-S.K. designed the study; A.P., H.G., M.-S.K. managed the study; D.G., C.L.K., C.A.I.-D., K.R.M. collected human cells/tissues; M.-S.K., R.C., D.G. developed the pipeline of experiment and analysis; D.G., M.-S.K., S.M.P., K.M., R.C., S.R., J.Z., X.W., P.G.S., M.S.Z., T.-C.H. prepared peptide samples for LC-MS/MS; M.-S.K., R.S.N., S.M.P., R.C., D.S.K., S.R., G.J.S. performed LC-MS/MS; M.-S.K., S.M.P., S.P., S.S.M., C.J.M., J.A. and A.K.M. processed MS data and managed data; A.K.M., S.S.M., B.G., A.H.P., Y.S., M.-S.K. performed comparison analysis with PeptideAtlas, neXtProt and GPMDB; R.I., S.Jai., G.D.B. performed interaction and complex analysis; M.-S.K., S.M.P., S.S.M., P.K., A.K.M., N.A.S., R.S.N., L.B., L.D.N.S., D.S.K., V.N., A.R., T.S., M.K., S.K.Sr., G.D., A.Mar., R.R., S.C., K.K.D., A.S., S.D.Y., S.Jay., P.R., A.H.P., B.G., J.S., N.S., R.G., G.J.S., A.A.K., S.A., D.F., T.S.K.P., H.G., A.P. performed proteogenomic analysis; A.C., H.L., R.S., J.T.S., K.K.M., S.S., A.Mah., S.K.Sh., P.S., S.D.L., C.G.D., A.Mai., M.K.H., R.H.H., C.L.K., C.A.I.-D. assisted with analysis of the data; M.-S.K., S.M.P., T.-C.H., P.L.-R. performed western blot experiments; M.-S.K., J.K.T., A.K.M., B.M., S.P., S.M.P. designed the Human Proteome Map web portal; M.-S.K., A.K.M., J.K.T. generated selected reaction monitoring (SRM) database; M.-S.K., K.M., G.D., S.M.P., S.S.M. illustrated figures with help of other authors; A.P., M.-S.K., H.G. wrote the manuscript with inputs from other authors.

Author Information The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD000561. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.G. (harsha@ibioinformatics.org) or A.P. (pandey@jhmi.edu).

ARTICLE

doi:10.1038/nature13319

Mass-spectrometry-based draft of the human proteome

Mathias Wilhelm[1,2]*, Judith Schlegl[2]*, Hannes Hahne[1]*, Amin Moghaddas Gholami[1]*, Marcus Lieberenz[2], Mikhail M. Savitski[3], Emanuel Ziegler[2], Lars Butzmann[2], Siegfried Gessulat[2], Harald Marx[1], Toby Mathieson[3], Simone Lemeer[1], Karsten Schnatbaum[4], Ulf Reimer[4], Holger Wenschuh[4], Martin Mollenhauer[5], Julia Slotta-Huspenina[5], Joos-Hendrik Boese[2], Marcus Bantscheff[3], Anja Gerstmair[2], Franz Faerber[2] & Bernhard Kuster[1,6]

Proteomes are characterized by large protein-abundance differences, cell-type- and time-dependent expression patterns and post-translational modifications, all of which carry biological information that is not accessible by genomics or transcriptomics. Here we present a mass-spectrometry-based draft of the human proteome and a public, high-performance, in-memory database for real-time analysis of terabytes of big data, called ProteomicsDB. The information assembled from human tissues, cell lines and body fluids enabled estimation of the size of the protein-coding genome, and identified organ-specific proteins and a large number of translated lincRNAs (long intergenic non-coding RNAs). Analysis of messenger RNA and protein-expression profiles of human tissues revealed conserved control of protein abundance, and integration of drug-sensitivity data enabled the identification of proteins predicting resistance or sensitivity. The proteome profiles also hold considerable promise for analysing the composition and stoichiometry of protein complexes. ProteomicsDB thus enables navigation of proteomes, provides biological insight and fosters the development of proteomic technology.

The large-scale interrogation of biological systems by mass-spectrometry-based proteomics provides insights into protein abundance, cell-type and time-dependent expression patterns, post-translational modifications (PTMs) and protein–protein interactions, all of which carry biological information that is best investigated at the protein level. Perhaps surprisingly, it is still not clear which of the 19,629 human genes annotated in Swiss-Prot[1] (20,493 in UniProt) are translated into proteins. Therefore, major efforts are underway to identify these genes, including the Human Proteome Project (HPP), which aims to broadly characterize the human proteome, the Human Protein Atlas project (HPA), which seeks to generate antibodies for all human proteins, and the ProteomeXchange consortium, which facilitates the gathering and sharing of proteomic data[2–4]. Despite the fact that a plethora of individual human proteomic studies exist, only a few systematic efforts to assemble and characterize human proteomes have been reported so far[5,6]. In part this is because most proteomic data do not reside in public repositories, proteomic data annotation is often sketchy, and the data generation and processing platforms are of varying capability, performance and maturity. Importantly, there is also a notable challenge in making 'big data' (that is, large amounts of data that are difficult or impossible to process using traditional technology and algorithms) more widely accessible to the scientific community, because the development of scalable analysis tools is only in its infancy.

Assembly of the proteome in ProteomicsDB

Here we present a draft of the human proteome assembled using data from 16,857 liquid chromatography tandem-mass-spectrometry (LC-MS/MS) experiments involving human tissues, cell lines, body fluids, as well as data from PTM studies and affinity purifications. We also present the analysis of the assembled data, in ProteomicsDB, an in-memory database designed for the real-time analysis of big data (https://www.proteomicsdb.org). For this study (Fig. 1a), we combined data available from repositories and otherwise contributed by colleagues (60% of total) with published as well as new data from the authors' laboratories (40% of total; Supplementary Table 1 and Supplementary Information). To maximize proteome coverage, we reprocessed all experiments using MaxQuant[7] and Mascot[8], and the resulting 1.1-billion peptide spectrum matches (PSMs) were imported into ProteomicsDB. The database (Fig. 1b) comprises a public repository, a web interface featuring several data views and analysis tools, and an application programming interface (API). At the heart of ProteomicsDB is an 'in-memory' computational resource commanding 2 terabytes (TB) of random access memory (RAM) and 160 central processor units (CPUs), which enables the storage of all data in the main memory, all of the time. This makes computational tasks very efficient, illustrated by the capability to display and annotate any of the approximately 71-million currently identified peptide-mass spectra in real time (Extended Data Fig. 1). Controlling the quality of peptide and protein identifications is important but exactly how this is best accomplished is still debated in the community[9,10]. For the current assembly of the proteome, we relied on high resolution mass-spectrometry data to keep false identifications low. We applied a two-step filtering process, first by controlling the false discovery rate (FDR) at 1% for PSMs generated by each LC-MS/MS experiment using a global target-decoy approach[11]. Peptide identifications then had to pass a length-dependent Mascot or Andromeda score threshold of 5% local FDR on the total aggregated data and we categorically rejected all peptides shorter than seven amino acids (Extended Data Figs 1 and 2, and Supplementary Information). Comparison to 27 published studies shows that this scheme is in line with the often-used 1% protein FDR criterion (Supplementary Table 1) and avoids the unsolved issue of artificially high protein FDRs when analysing large data sets[12] (Extended Data Fig. 2 and Supplementary Information).

Proteomic annotation of the genome

At the time of writing, ProteomicsDB held protein evidence for 18,097 of the 19,629 human genes annotated in Swiss-Prot (92%) as well as

[1]Chair of Proteomics and Bioanalytics, Technische Universität München, Emil-Erlenmeyer Forum 5, 85354 Freising, Germany. [2]SAP AG, Dietmar-Hopp-Allee 16, 69190 Walldorf, Germany. [3]Cellzome GmbH, Meyerhofstraße 1, 69117 Heidelberg, Germany. [4]JPT Peptide Technologies GmbH, Volmerstraße 5, 12489 Berlin, Germany. [5]Institute of Pathology, Technische Universität München, Trogerstraße 18, 81675 München, Germany. [6]Center for Integrated Protein Science Munich, Germany.
*These authors contributed equally to this work.

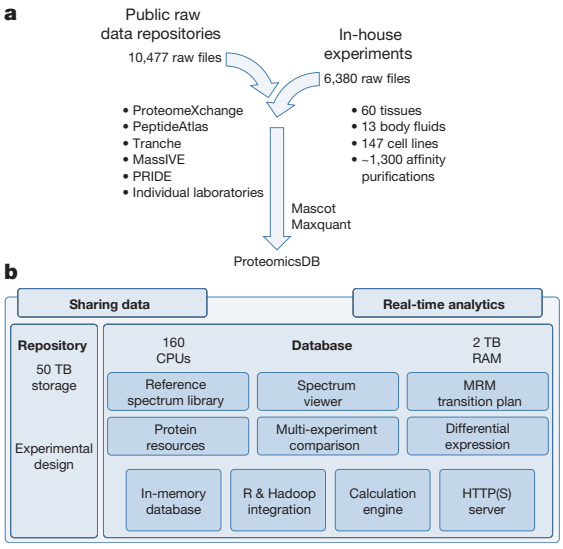


Figure 1 | Strategy for the assembly of the human proteome. **a**, Experimental workflow for the identification and quantification of proteins. **b**, Structure and features of ProteomicsDB. ProteomicsDB consists of a repository part for raw-data storage and an in-memory database designed for the storage, analysis and visualization of proteomic data sets. Fast computation on large data sets is backed by 160 CPUs and 2 TB of RAM.

19,376 out of 86,771 protein isoforms listed in UniProt (22%; Supplementary Table 1). Chromosomes were evenly covered with the notable exceptions of chromosome 21 and the Y chromosome (Fig. 2a). The former contains many proteins with few mass-spectrometry-compatible tryptic peptides. As 257 human proteins (not counting isoforms) do not produce any such peptides, this renders trypsin—as the most frequently used protease in proteomics—ineffective. As a result, alternative proteases or top down sequencing approaches will have a part to play in the eventual completion of the human proteome (Extended Data Fig. 3a)[13,14]. To facilitate this, ProteomicsDB provides a tool predicting the best protease or combinations thereof for any protein which can also be valuable when systematically mapping PTMs.

We next attempted to estimate the size of the protein-coding genome based on UniProt protein evidence categories. ProteomicsDB currently covers 97% of the 13,378 genes with annotated evidence on protein and 84% (of 5,531) with evidence on transcript level. The overlap with proteins detected by antibodies in the HPA project is 93% (of 15,156 HPA proteins) providing independent evidence that these genes exist as proteins. Conversely, proteomic coverage of genes inferred from homology (52% of 159), genes marked as predicted (64% of 72) or uncertain (56% of 489), was considerably lower, suggesting that the protein-coding human genome may be several hundred genes smaller than anticipated previously. Still, we were able to validate the identification of 36 of the uncertain genes (out of 44 tried)[15] using reference spectra from synthetic peptides (Supplementary Table 2). Among the identified uncertain genes were three lincRNAs (Extended Data Fig. 3). This unexpected result prompted us to search approximately 9-million tandem-mass-spectrometry spectra from tissues and cell lines against 13,564 lincRNA sequences from Ensembl and 21,487 lincRNAs and TUCPs (transcripts of uncertain coding potential) from the Broad Institute[16]. This returned 430 high-quality peptides (no homology to UniProt sequences) from 404 lincRNAs and TUCPs (Supplementary Table 3). There was no apparent bias in chromosomal location or biological source, and the abundance distribution of translated lincRNA peptides was broadly similar to that of peptides from ordinary proteins (Extended Data Fig. 3). To our knowledge, this is the largest number of lincRNA and TUCP translation products with direct peptide evidence reported to date[17], arguing that translation of such transcripts is more common than anticipated previously[18–20]. The biological significance of translated lincRNAs and TUCPs is not clear at present. These may constitute proteins 'in evolution' representing hitherto undiscovered biology[21] or arise by stochastic chance marking such proteins as 'biological noise'.

Core proteome and missing proteome

Aggregation of the data used for building the draft proteome shows that proteome coverage rapidly saturates at approximately 16,000 –17,000 proteins, which is similar to transcriptome coverage obtained by RNA sequencing (RNA-seq). Addition of human-tissue and body-fluid data each led to small but noticeable contributions not provided by cell lines. The same is true when adding PTM or affinity data to shotgun proteomic data (Extended Data Fig. 4). When comparing five of the largest data sets in ProteomicsDB[22–25], the existence of a human core proteome[26] of approximately 10,000–12,000 ubiquitously expressed proteins can be postulated, the primary function of which is the general control and maintenance of cells (Extended Data Fig. 4 and Supplementary Table 4). The low abundance range of the core proteome is enriched in proteins with regulatory functions. The observed proteome saturation implies that adding more shotgun data will not considerably increase coverage, although it would increase confidence in individual proteins. Instead, it is likely that the 'missing proteome' (Fig. 2b and Supplementary Table 4) will have to be identified by more focused experimentation. It is also possible that a considerable part of the missing proteome constitutes (pseudo)genes that are no longer expressed. G-protein-coupled receptors (GPCRs) are underrepresented in ProteomicsDB and the respective transcripts are also notoriously absent in RNA-seq data[27]. Earlier work suggests that more than half of the 853 human GPCRs have lost their function over the course of human evolution and may be considered obsolete[28]. Similarly, a large number of functionally uncategorized proteins are annotated pseudogenes, potentially further reducing the number of (actual) protein-coding genes. Cytokines may be underrepresented because of experimental issues as small, secreted proteins can

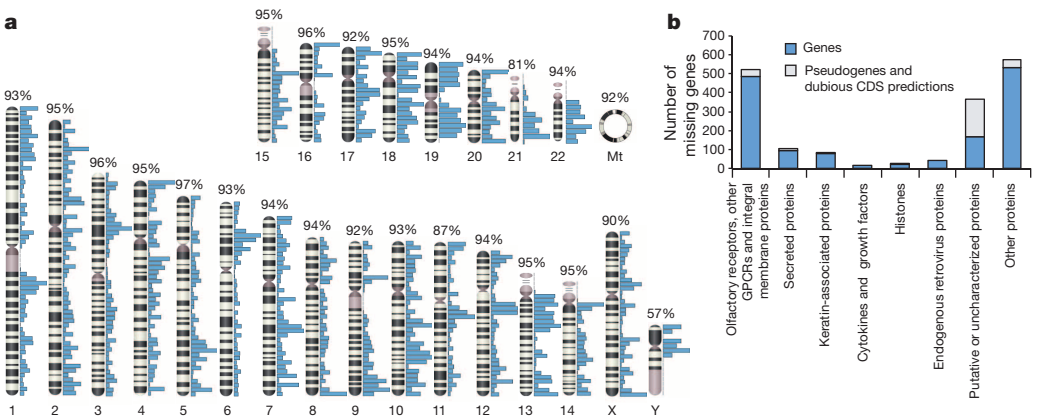


Figure 2 | Characterization of the human proteome. **a**, Chromosomal coverage of the 18,097 proteins identified in this study exceeds 90% in all but three cases. Blue bars indicate the density of proteins in a particular chromosomal region. **b**, Gene ontology analysis of the 'missing' proteome' identifies GPCRs, secreted and keratin-associated proteins as the major protein classes underrepresented in proteomic experiments. CDS, coding sequence; Mt, mitochondrial DNA.

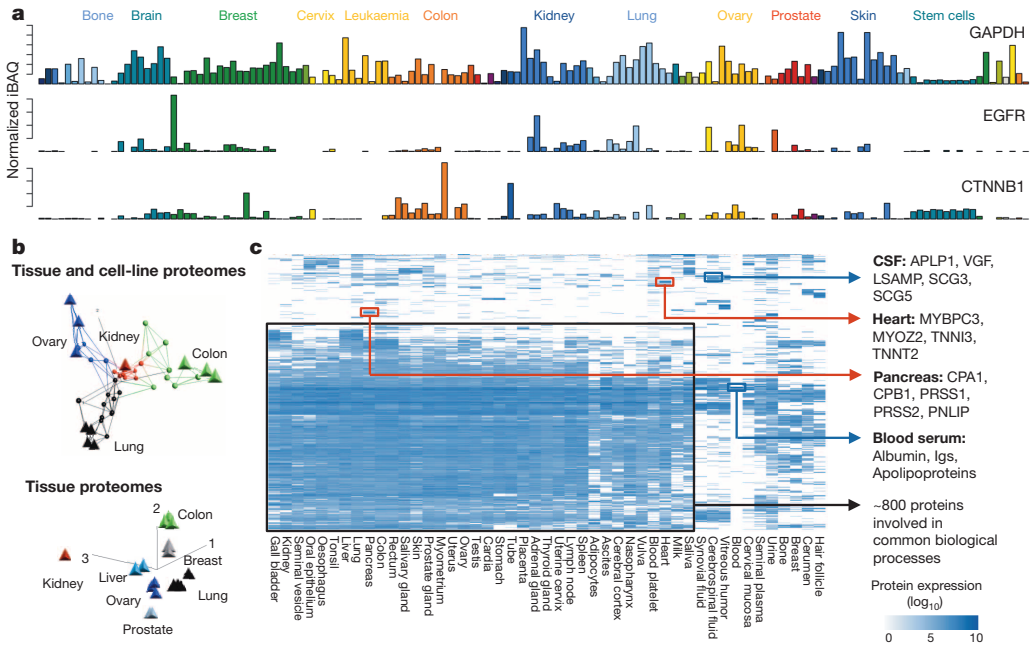


Figure 3 | Global protein expression analysis. **a**, Protein expression in different tissues and cell lines, showing that levels of housekeeping (GAPDH), signalling (EGFR) and tumour-associated (CTNNB1) proteins can vary substantially between tissues (grouped by colour). **b**, PCA showing that cell lines (circles) retain protein-expression characteristics of their respective primary tissue (triangles) and that proteomes of different organs are more diverse. **c**, Hierarchical clustering of the 100 most highly expressed proteins from each of 47 tissues and body fluids. Despite the presence of a large group of common proteins, clusters of organ and fluid-selective proteins with respective biological functions can readily be identified.

still be difficult to obtain from the supernatants of cells, the intercellular space of tissues or from body fluids. To fill the remaining gaps in the human proteome, ProteomicsDB provides a facility to engage the community by 'adopting' a missing protein; that is, to provide mass-spectrometric evidence for its existence. In addition, we have synthesized and identified 435 peptides for all 273 cytokines as well as 3,539 further peptides for proteins not yet well covered and have made their tandem-mass-spectrometry spectra available in ProteomicsDB so that any identification of such proteins in the future may be validated using the synthetic reference standard (Supplementary Table 2).

Functional proteome–expression analysis

We have generated proteome profiles of 27 human tissues and body fluids (human body map) complemented with publically available data (Supplementary Tables 1 and 5) to begin to analyse human proteomes in functional terms. To normalize the disparate data sets in Proteomics DB, we found the intensity-based absolute-protein-quantification method (iBAQ) to be appropriate (Extended Data Fig. 5 and Supplementary Information)[29–31]. A simple common task is to compare the expression level of a single protein across many biological sources (Fig. 3a). Although housekeeping proteins such as GAPDH (glyceraldehyde-3-phosphate dehydrogenase) show high (and sometimes extreme) expression throughout biological sources, high levels of the proto-oncogene EGFR (epidermal growth factor receptor) are mostly confined to cancerous tissue; for example, breast cancer tissue. Similarly, β-catenin, a member of the Wnt pathway, is highly expressed in colon cancer cells, where the protein participates in the development of the malignancy. Principle component analysis (PCA) of protein abundances in 42 proteomes shows that protein expression in a particular tissue and its corresponding cell lines is broadly similar and that there are more substantial differences between tissues of different organs (Fig. 3b). This result is important for the interpretation of data presented below and also contributes to the ongoing discussion regarding the suitability of cell lines as model systems for studying human biology. A comparative analysis of the 100 most highly expressed proteins in each of 47 human organs and body fluids (Fig. 3c) revealed that approximately 70% of these proteins are found in all organs and body fluids but show expression differences of up to five orders of magnitude (Supplementary Table 4). Interestingly, even the most highly abundant proteins in a tissue or fluid often point to molecular processes associated with the respective biological specialization; myofibrillar proteins, including troponins, are abundant in the heart, proteases in the pancreas and neuronal proteins in cerebrospinal fluid.

Similar observations can be made when investigating proteins forming functional classes such as protein kinases or transcription factors (Fig. 4 and Supplementary Table 4). Akin to core proteomes, some of the 349 detected kinases and 557 transcription factors are broadly expressed, but others seem to be confined to few organs where they drive more specific processes. For example, the kinases HCK, ZAP70, LCK, JAK3, TXK and FGR are found in a tight cluster of kinases in the spleen and all have important roles in the biology of immune cells. This is 'mirrored' by transcription factors in the same cluster with strong ties to immunity, including the NF-κB system (REL, PRKCH, NFKBIE) and Toll-like receptor signalling (SIGIRR, IRF5, ARRB2, NLRC4). It is noteworthy that many of the proteins in the spleen cluster are also highly

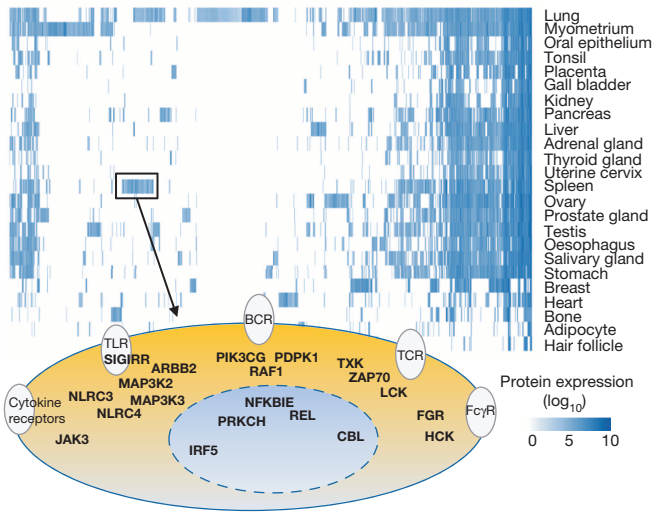


Figure 4 | Functional protein expression analysis. Quantitative expression analysis of 906 kinases and transcription factors across 24 tissues (top panel) identifies organ-selective signatures indicative of the underlying biology. The highlighted cluster in spleen contains the kinases LCK, ZAP70 and JAK and the transcription factors SIGIRR, NFKBIE and NLRC3 with strong links to the immune system (bottom panel). Yellow oval represents a cell; blue oval represents the nucleus.

expressed in the lung, a primary entry point for human pathogens. The number of proteins that are exclusively or preferentially detected in a particular organ is surprisingly small, and gene ontology analysis invariably highlights organ-specific biology (Extended Data Fig. 6). For example, adipocytes are rich in proteins involved in lipid storage, platelets in growth factors, and placenta in proteins relating to hormonal regulation and pregnancy (Supplementary Table 5). The above shows that even disparate, though high-quality proteomic data can be used to construct protein expression maps across an entire complex organism. A recent report has shown that this is feasible in mice[32] but to our knowledge, organism-wide proteome-expression profiling has not been described in humans before. In addition, the identification of a considerable number of proteins with no ascribed function but exclusive (or high) expression in particular organs implies a functional role. The contextual information provided in ProteomicsDB may thus provide guidance for the eventual identification of the biological role of these orphan proteins.

Integration and utility of proteomes

Many further uses of protein-expression profiling can be envisaged, of which we can only outline a few here. We have compared messenger RNA (RNA-seq)[27] and protein (iBAQ, this study) expression profiles for 12 human tissues (Extended Data Fig. 7, Supplementary Table 6 and Supplementary Information) and clear correlations are observed in all cases but the Spearman's rank correlation coefficients are rather moderate and somewhat poorer than those previously reported for cell lines. This is likely to be due to the fact that tissues generally comprise a mixture of cell types, connective tissue and blood. Both mRNA and protein levels vary greatly between tissues as one might expect; however, the ratio of protein and mRNA levels is remarkably conserved between tissues for any given protein (Fig. 5a, top panel)[33]. It has been shown previously[31] that the translation rate constant is one dominant factor determining protein abundance in cell lines. Using the ratio of protein to mRNA levels as a proxy for translation rates, our data show

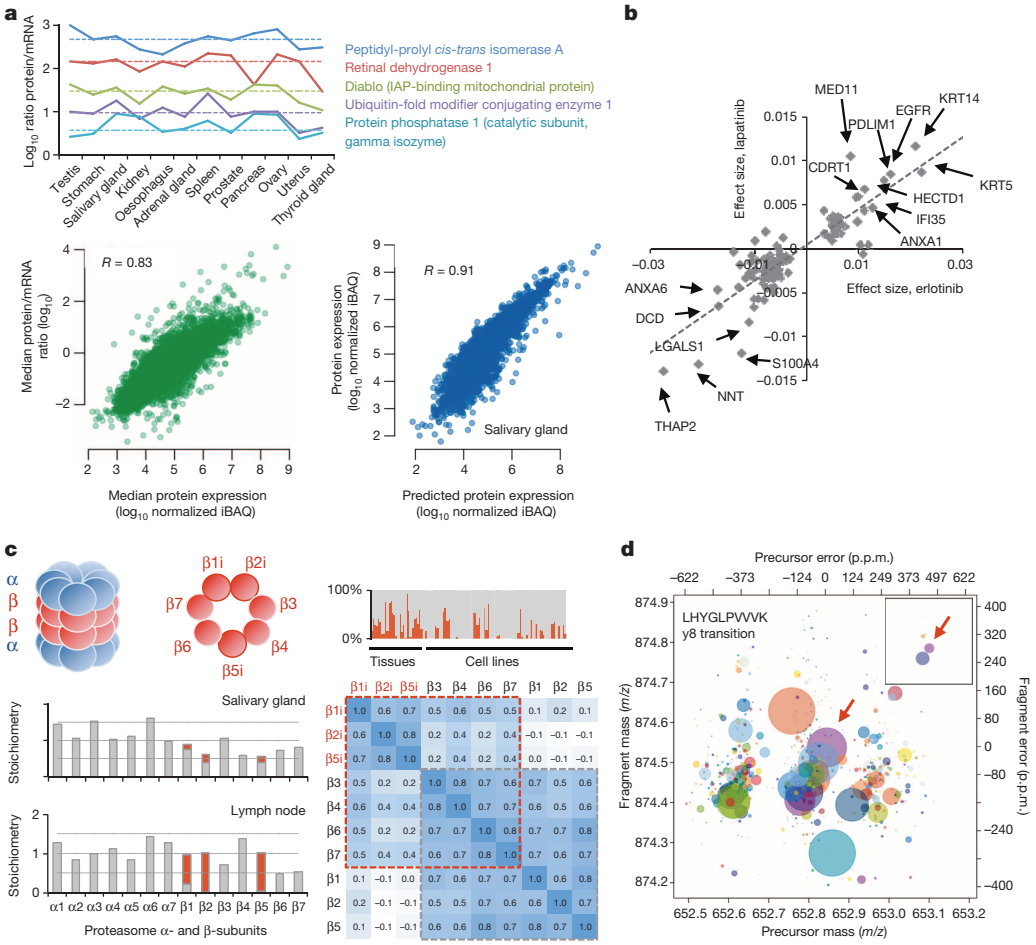


Figure 5 | Integration and utility of large proteomic data collections. **a**, Analysis of mRNA and protein levels across 12 organs shows that the protein/mRNA ratio is largely conserved (top panel). The median translation rates of all transcripts across all tissues correlate well with protein abundance (bottom-left panel), leading to the ability to predict individual protein levels from the respective mRNA levels (bottom-right panel). **b**, Elastic net analysis for the identification of drug sensitivity (positive-effect-size) or resistance (negative-effect-size) markers against the EGFR kinase inhibitors erlotinib and lapatinib in cancer cell lines. **c**, Analysis of the composition and stoichiometry of the proteasome. Top-left panel, schematic structure of the 'constitutive' proteasome and the 'immunoproteasome' (marked by the suffix 'i'). Middle-left and bottom-left panels, stoichiometry derived by iBAQ of the constitutive proteasome (grey) and the immunoproteasome (red) in the salivary gland and the lymph node. Top-right panel, expression analysis of the β1 subunit across more than 100 tissue and cell-line proteomes reveals that many cells express both forms of the proteasome. Bottom-right panel, expression correlation analysis of all β subunits across the said tissues and cell lines showing strong co-expression of the β1i, β2i and β5i subunits as well as all other β-subunits but no correlation with the expression of the corresponding β1, β2 and β5 subunits. **d**, ProteomicsDB enables the computation of molecular interferences in selected reaction-monitoring experiments (SRM) from experimental data. The transition of the target peptide LHYGLPVVVK (y8 fragment ion, β-catenin) is marked with an arrow. All other circles in the plot are interfering SRM transitions of other peptides found in ProteomicsDB that fall within the same mass tolerance of the experiment (here, 0.7 Da). The size of each circle indicates the severity of the interference. The inset shows that interference can be substantially reduced by the use of high-resolution fragment-ion data (here, 0.04 Da) and confining the analysis to the tissue from which a sample is derived (here, a colon sample).

that this is also true for human tissues and that the ratio is similar in every tissue (Fig. 5a, bottom left panel). It therefore appears that the translation rate is a fundamental, encoded (constant) characteristic of a transcript, suggesting that the actual amount of protein in a given cell is primarily controlled by regulating mRNA levels. Having learned the protein/mRNA ratio for every protein and transcript, it now becomes possible to predict protein abundance in any given tissue with good accuracy from the measured mRNA abundance (Fig. 5a, bottom right panel, and Extended Data Fig. 7).

We have shown previously that protein expression can be correlated to drug sensitivity[24]. Here we used drug-sensitivity data provided by the cancer cell line encyclopedia[34] (CCLE) to discover sensitivity and resistance markers for 24 drugs in 35 human cancer cell lines (Supplementary Table 7). For the EGFR kinase inhibitors erlotinib and lapatinib the primary target (EGFR) as well as annexin A1 (ANXA1, a direct EGFR substrate), and EGFR interacting proteins at stress fibres (PDLIM, KRT5, KRT14) all indicate drug sensitivity, whereas high expression of ANXA6 or S100A4 renders cells less responsive (Fig. 5b and Extended Data Fig. 8). Assuringly, knockdown of ANXA6 in BT549 cells has been shown to sensitize cells to lapatinib[35] and addition of S100A4 to cells in culture has been shown to stimulate EGFR and to promote metastasis[36]. High expression of S100 proteins is often associated with resistance against kinase inhibitors (Supplementary Table 7), suggesting that this may constitute a general molecular resistance mechanism. Similar effects can be postulated for other proteins (Supplementary Notes) and in light of a recent report showing increased phosphorylation of HECTD1 on EGF treatment[37], it is tempting to speculate that a HECTD1–CDRT1 E3 ubiquitin ligase–orphan F-box protein complex may be involved in regulating the stability of EGFR via the ubiquitin–proteasome system.

The composition and stoichiometry of protein complexes is typically analysed by affinity purification coupled to mass-spectrometry-based protein analysis and it emerges that protein expression profiling may also have potential for this purpose[38]. We found that stoichiometries measured by iBAQ for the nuclear pore complex agreed well with a prior study using absolute protein quantification by spiked peptide standards (Extended Data Fig. 9 and Supplementary Table 8)[39]. Using the proteasome as an example, we explored its composition and stoichiometry heterogeneity across cell lines and tissues (Fig. 5c). The constitutive core proteasome consists of 2×7 non-catalytic α- and 2×7 catalytic β-subunits but, for example, an 'immunoproteasome' has been identified in which the β1, 2 and 5 subunits are replaced by homologous proteins (β1i, β2i and β5i) in immune cells[40,41]. Our analysis shows that the proteasome in the salivary gland is primarily of the constitutive type and that lymph nodes almost exclusively contain the immunoproteasome (Fig. 5c, left panel). The same analysis across more than 100 cell-line and tissue samples (Fig. 5c, right panel) reveals that the immunoproteasome is surprisingly widely expressed, including in tissues for which no primary immunological function would be expected. In addition, the data imply that the molecular composition and stoichiometry of proteasomes is heterogeneous and cell-type-dependent. Correlation analysis of the expression of all β-subunits (Fig. 5c, bottom right panel) strongly suggests that the β1, 2 and 5 subunits and their respective immunoproteasome counterparts are expressed independently (no correlation). In contrast, it seems that the remaining subunits (β3, 4, 6, 7) are co-expressed with either group.

Proteomic data collections can be valuable data mines for post-translational modification analysis or developing proteome technology. ProteomicsDB currently contains 81,721 unique phosphorylated peptides representing 11,025 human genes, demonstrating that more than half of all human proteins are substrates of kinases. Similarly, there are 29,031 unique ubiquitinylated peptides from 5,769 proteins representing substrates of ubiquitin ligases as well as 16,693 acetylated peptides from 7,098 proteins that are substrates of acetylases. Our analysis also detected amino-terminal peptides for 7,977 proteins and carboxy-terminal peptides for 6,778 proteins confirming a large number of translation start and stop sites (Extended Data Fig. 10a). We expect that the PTM branch of ProteomicsDB will grow rapidly over time and help to build a future version of the human proteome that provides more direct links between protein expression and activity.

So-called 'proteotypic' peptides[42] have proven useful as quantification standards in targeted proteomic measurements, which are increasingly employed to develop clinical biomarker assays[43]. ProteomicsDB enabled us to determine the proteotypicity of ~approximately 500,000 peptides and to expand the concept to chemically labelled peptides (Extended Data Fig. 10b, Supplementary Table 9 and Supplementary Information). The 71-million peptide-precursor and 18-billion peptide-fragment ion measurements enables the computational assessment of the specificity of targeted measurements ahead of the actual experiment. Exemplified by the peptide LHYGLPVVVK of the proto-oncogene β-catenin (Fig. 5d and Extended Data Fig. 10c), mining of ProteomicsDB revealed a large number of potentially interfering peptides that may distort the quantification of the target peptide. Interference can be substantially reduced by high-resolution instruments[44] and by limiting the allowed interferences to the tissue in question (Fig. 5d, inset). We anticipate that the combination of experimental proteotypicity, interference estimation and high-resolution instrumentation will provide for more robust targeted proteomic assays in the future.

Discussion

Here we have shown that an extensive draft of the human proteome can be assembled from disparate but high-quality proteomic data. We have outlined some of the many applications that can be envisaged for its use and some of the biological insights that may be generated by mining the proteome. Similar to the evolution of the human genome projects, the eventual completion of the human proteome will take further time and effort but will also lead to substantial improvements in technology, which are still needed. One issue to address is proteome coverage and resolution. While DNA and RNA sequencing technologies have attained single-nucleotide resolution, the amino-acid coverage of proteins is still limited, which currently impairs our ability to detect protein variants, such as differential splice products, PTMs, mutations or isoforms in a systematic fashion. A related challenge is to improve the ability to sample a proteome comprehensively; that is, 'all proteins, all the time'. Another important area of future research concerns overcoming the uncertainties associated with peptide and protein identification by sequence-database searching[45]. ProteomicsDB and similar resources have a part to play in these challenges as the data assembled will enable the development of computational tools and laboratory reagents facilitating proteome-wide discovery experiments, multiplexed quantitative protein assays, as well as general exploration of the human proteome.

METHODS SUMMARY

Proteomic data were downloaded from public repositories, contributed by individual laboratories and specifically generated for this study by the authors' laboratories. For the specifically generated data, human tissue specimens were obtained from the Biobank of the Technische Universität München following approval of the study by the local ethics committee. Samples were collected within the first 30 min after resection, macroscopically resected by an experienced pathologist, snap frozen and stored in liquid nitrogen until use. Body fluids requiring no invasive procedures were provided by volunteers. Proteins were extracted under denaturing conditions and either separated by LDS-PAGE followed by in-gel protease digestion or digested in solution in the presence of chaotropic agents. Synthetic peptides were produced by solid-phase chemistry following the standard Fmoc strategy and used without purification. Peptides were separated by ultra-high-pressure liquid chromatography and analysed on Orbitrap mass spectrometers using either resonance-type or beam-type collision-induced dissociation. For peptide identification, tandem mass spectra were processed in parallel using Mascot Distiller and MaxQuant with Andromeda[7], and searched against UniProt and/or a custom build fasta-formatted sequence file containing lincRNA sequences. Search results and tandem mass spectra were imported into ProteomicsDB (https://www.proteomicsdb.org) and filtered at 1% PSM FDR and 5% local peptide-length-dependent FDR. For bioinformatic analysis, data were extracted from ProteomicsDB using HANA Studio and further processed using custom python scripts and statistics programme R. Gene ontology analysis was

performed using David (http://david.abcc.ncifcrf.gov) and REViGO (http://revigo.irb.hr/). See Supplementary Information for details.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 November 2013; accepted 11 April 2014.

1. UniProt. C. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **41,** D43–D47 (2013).
2. Paik, Y. K. *et al.* The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nature Biotechnol.* **30,** 221–223 (2012).
3. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nature Biotechnol.* **28,** 1248–1250 (2010).
4. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnol.* **32,** 223–226 (2014).
5. Farrah, T. *et al.* State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *J. Proteome Res.* **13,** 60–75 (2014).
6. Wang, M. *et al.* PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics* **11,** 492–500 (2012).
7. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnol.* **26,** 1367–1372 (2008).
8. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).
9. Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **22,** 1111–1120 (2011).
10. Higdon, R. *et al.* IPM: An integrated protein model for false discovery rate estimation and identification in high-throughput proteomics. *J. Proteomics* **75,** 116–121 (2011).
11. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnol.* **24,** 1285–1292 (2006).
12. Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **8,** 2405–2417 (2009).
13. Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7,** 548 (2011).
14. Tran, J. C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480,** 254–258 (2011).
15. Lane, L. *et al.* Metrics for the Human Proteome Project 2013–2014 and strategies for finding missing proteins. *J. Proteome Res.* **13,** 15–20 (2014).
16. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25,** 1915–1927 (2011).
17. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
18. Bánfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22,** 1646–1657 (2012).
19. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154,** 240–251 (2013).
20. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).
21. Flintoft, L. Non-coding RNA: Ribosomes, but no translation, for lincRNAs. *Nature Rev. Genet.* **14,** 520 (2013).
22. Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11,** M111.014050 (2012).
23. Mertins, P. *et al.* Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature Methods* **10,** 634–637 (2013).
24. Moghaddas Gholami, A. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell Rep.* **4,** 609–620 (2013).
25. Shiromizu, T. *et al.* Identification of missing proteins in the neXtProt database and unregistered phosphopeptides in the PhosphoSitePlus database as part of the Chromosome-centric Human Proteome Project. *J. Proteome Res.* **12,** 2414–2421 (2013).
26. Schirle, M., Heurtier, M. A. & Kuster, B. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2,** 1297–1305 (2003).
27. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13,** 397–406 (2014).
28. Hughes, G. M., Teeling, E. C. & Higgins, D. G. Loss of olfactory receptor function in hominin evolution. *PLoS ONE* **9,** e84714 (2014).
29. Ahrné, E., Molzahn, L., Glatter, T. & Schmidt, A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* **13,** 2567–2578 (2013).
30. Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7,** 549 (2011).
31. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).
32. Geiger, T. *et al.* Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Mol. Cell. Proteomics* **12,** 1709–1722 (2013).
33. Low, T. Y. *et al.* Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep.* **5,** 1469–1478 (2013).
34. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
35. Koumangoye, R. B. *et al.* Reduced annexin A6 expression promotes the degradation of activated epidermal growth factor receptor and sensitizes invasive breast cancer cells to EGFR-targeted tyrosine kinase inhibitors. *Mol. Cancer* **12,** 167 (2013).
36. Klingelhöfer, J. *et al.* Epidermal growth factor receptor ligands as new extracellular targets for the metastasis-promoting S100A4 protein. *FEBS J.* **276,** 5936–5948 (2009).
37. Argenzio, E. *et al.* Proteomic snapshot of the EGF-induced ubiquitin network. *Mol. Syst. Biol.* **7,** 462 (2011).
38. Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150,** 1068–1081 (2012).
39. Ori, A. *et al.* Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. *Mol. Syst. Biol.* **9,** 648 (2013).
40. Hisamatsu, H. *et al.* Newly identified pair of proteasomal subunits regulated reciprocally by interferon gamma. *J. Exp. Med.* **183,** 1807–1816 (1996).
41. Nandi, D., Jiang, H. & Monaco, J. J. Identification of MECL-1 (LMP-10) as the third IFN-gamma-inducible proteasome subunit. *J. Immunol.* **156,** 2361–2364 (1996).
42. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnol.* **25,** 125–131 (2007).
43. Domon, B. Considerations on selected reaction monitoring experiments: implications for the selectivity and accuracy of measurements. *Proteomics Clin. Appl.* **6,** 609–614 (2012).
44. Gallien, S. *et al.* Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol. Cell. Proteomics* **11,** 1709–1723 10.1074/mcp.O112.019802 (2012).
45. Marx, H. *et al.* A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nature Biotechnol.* **31,** 557–564 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors wish to thank all originators of the mass-spectrometry-data used in this study for making their data available. We are grateful to P. Mallick, J. Cottrell and M. Schirle for conceptual discussions, to F. Pachl, S. Heinzlmeir, S. Klaeger, S. Maier, D. Helm, B. Ferreia, M. Frejno, H. Koch, M. Mundt, J. Zecha, D. Zolg, E. Gillmeier, B. Ruprecht, K. Kramer, G. Medard and X. Ku of TUM for the annotation of experiments, and to Y. Morad, A. Niadzelka, E. Kny, H. Cossmann, D. Schikora of SAP and V. Wichnalek, A. Klaus, M. Kroetz-Fahning, T. Schmidt of TUM for technical assistance.

Author Contributions M.W., J.S., M.L., E.Z., L.B., J.-H.B., S.G., A.G., H.H., A.M.G. and B.K. designed ProteomicsDB. H.H., K.S., U.R., M.M. and J.S.-H. performed experiments. M.W., H.H., A.M.G., M.M.S., H.M., T.M., S.L. and B.K. performed data analysis. H.W., M.B., F.F. and B.K. conceptualized the study. M.W., H.H., A.M.G. and B.K. wrote manuscript.

Author Information Mass-spectrometry data are available from ProteomicsDB (https://www.proteomicsdb.org) and ProteomeXchange (http://proteomecentral.proteomexchange.org; dataset identifier PXD000865). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.K. (kuster@tum.de).

ARTICLE

doi:10.1038/nature13271

Structural basis of the non-coding RNA RsmZ acting as a protein sponge

Olivier Duss[1], Erich Michel[1], Maxim Yulikov[2], Mario Schubert[1], Gunnar Jeschke[2] & Frédéric H.-T. Allain[1]

MicroRNA and protein sequestration by non-coding RNAs (ncRNAs) has recently generated much interest. In the bacterial Csr/Rsm system, which is considered to be the most general global post-transcriptional regulatory system responsible for bacterial virulence, ncRNAs such as CsrB or RsmZ activate translation initiation by sequestering homodimeric CsrA-type proteins from the ribosome-binding site of a subset of messenger RNAs. However, the mechanism of ncRNA-mediated protein sequestration is not understood at the molecular level. Here we show for *Pseudomonas fluorescens* that RsmE protein dimers assemble sequentially, specifically and cooperatively onto the ncRNA RsmZ within a narrow affinity range. This assembly yields two different native ribonucleoprotein structures. Using a powerful combination of nuclear magnetic resonance and electron paramagnetic resonance spectroscopy we elucidate these 70-kilodalton solution structures, thereby revealing the molecular mechanism of the sequestration process and how RsmE binding protects the ncRNA from RNase E degradation. Overall, our findings suggest that RsmZ is well-tuned to sequester, store and release RsmE and therefore can be viewed as an ideal protein 'sponge'.

In the last decade, ncRNAs have emerged in both eukaryotes and prokaryotes (in which they are termed small RNAs, or sRNAs) as important regulators in gene expression[1–3]. The bacterial Csr/Rsm system (for carbon storage regulator or repressor of secondary metabolism) is a well-characterized sRNA-based regulatory system that is present in about 75% of all bacterial species and affects over 20% of all mRNAs in *Escherichia coli*[3–6]. The homodimeric CsrA/RsmE proteins bind to the ribosome-binding site (RBS) of a subset of mRNAs and repress translation initiation[5–7]. Triggered by environmental signals, one or several sRNAs (for example CsrB or RsmZ) are transcribed, which can sequester the CsrA/RsmE proteins and therefore de-repress translation initiation. Both the sRNAs as well as the regulated mRNAs contain several GGA protein-binding motifs that bind the CsrA/RsmE proteins[5–8]. Although the RNA-binding surfaces of all the CsrA/RsmE homologues are highly conserved, the sRNAs are very diverse in length (100–500 nucleotides), predicted secondary structures, and numbers of GGA-binding motifs[5,6].

We previously elucidated the molecular basis of translation repression by solving the solution structure of RsmE bound to a stem–loop RNA encompassing the Shine-Dalgarno sequence of the *hcnA* mRNA of *P. fluorescens*[7]. However, no information is available on the mechanism of protein sequestration by any sRNA. Here we report a combined NMR- and electron paramagnetic resonance (EPR) spectroscopy-based approach that enabled us to visualize and rationalize the cooperative assembly of several RsmE proteins on the sRNA RsmZ from *P. fluorescens* and to obtain the 70-kilodalton (kDa) solution structure of the first 72 nucleotides of RsmZ, RsmZ(1–72), bound to three homodimeric RsmE proteins. These structural data provide detailed insight into the mechanism of translation activation by the sRNA RsmZ and also suggest how coupling with sRNA degradation allows the release of the RsmE protein.

Ordered RsmE–RsmZ RNP assembly

The 127-nucleotide sRNA RsmZ is composed of four stem–loops and a protective rho-independent terminator[9]. It contains eight GGA motifs, four being present in the loop of RNA hairpins and four located in single-stranded regions (Fig. 1a). In isolation, all eight GGA motifs of RsmZ bind RsmE with dissociation constants (K_d) ranging from 16 nM to 50 μM (Fig. 1a)[10]. Interestingly, only the GGA motif within stem–loop 2 (SL2) binds RsmE with low-nM K_d, whereas all other GGA motifs have a K_d in the μM-range. Considering that RsmE is tightly bound to the RBS of the *hcnA* mRNA[10] (K_d = 180 ± 40 nM, see Methods), efficient translation activation by RsmZ can only be understood by investigating the binding of RsmE to the entire RsmZ sRNA.

We performed native polyacrylamide RNA gels by incubating the RsmZ sRNA with increasing stoichiometric amounts of RsmE protein. Notably, the sRNA can accommodate sequentially up to five RsmE dimers (see Fig. 1b). To determine in which order the different GGA motifs are bound by RsmE, we followed NMR chemical shift perturbations by titrating increasing stoichiometric amounts of ^{2}H–^{15}N-labelled RsmE protein to the RNA RsmZ. By comparing the ^{1}H–^{15}N heteronuclear single-quantum coherence spectra (HSQC) of RsmE bound to isolated GGA motifs with the spectra of the RsmE dimer bound to RsmZ(1−89) at a 1:1 or 2:1 ratio, we could conclude that the first RsmE protein dimer binds simultaneously the GGA motifs of SL2 and SL3 and the second RsmE dimer the motifs of SL1 and SL4 (Fig. 1c).

A similar approach to monitor the additional RsmE binding sites failed owing to strong spectral overlap of the protein resonances. We then prepared segmentally labelled RNAs[11,12] containing only one GGA motif isotopically labelled in each case (GGA_{39-41}, GGA_{76-78} or GGA_{85-87}). When RsmE binds a GGA motif, the central guanosine imino proton forms an intermolecular hydrogen bond with the protein and becomes detectable by NMR spectroscopy. These experiments (Fig. 1d) revealed that the third RsmE dimer is simultaneously bound by the GGA motif in the single-stranded region between SL2 and SL3 (GGA_{39-41}), and by GGA_{85-87} in the single-stranded region between SL4 and the terminator (Fig. 1d, g). The fourth dimer is specifically bound by GGA_{76-78}. We then estimated the dissociation constants for the binding of the several RsmE proteins using the entire RsmZ sRNA by electrophoretic mobility shift assays (EMSA). We found that the first four dimers all have submicromolar dissociation constants (100–400 nM) whereas the binding of the fifth RsmE protein is not detectable using EMSA (>10 μM) (Fig. 1e, g). As all of the isolated binding sites except SL2 have micromolar

[1]Institute of Molecular Biology and Biophysics, ETH Zürich, CH-8093 Zürich, Switzerland. [2]Laboratory of Physical Chemistry, ETH Zürich, CH-8093 Zürich, Switzerland.

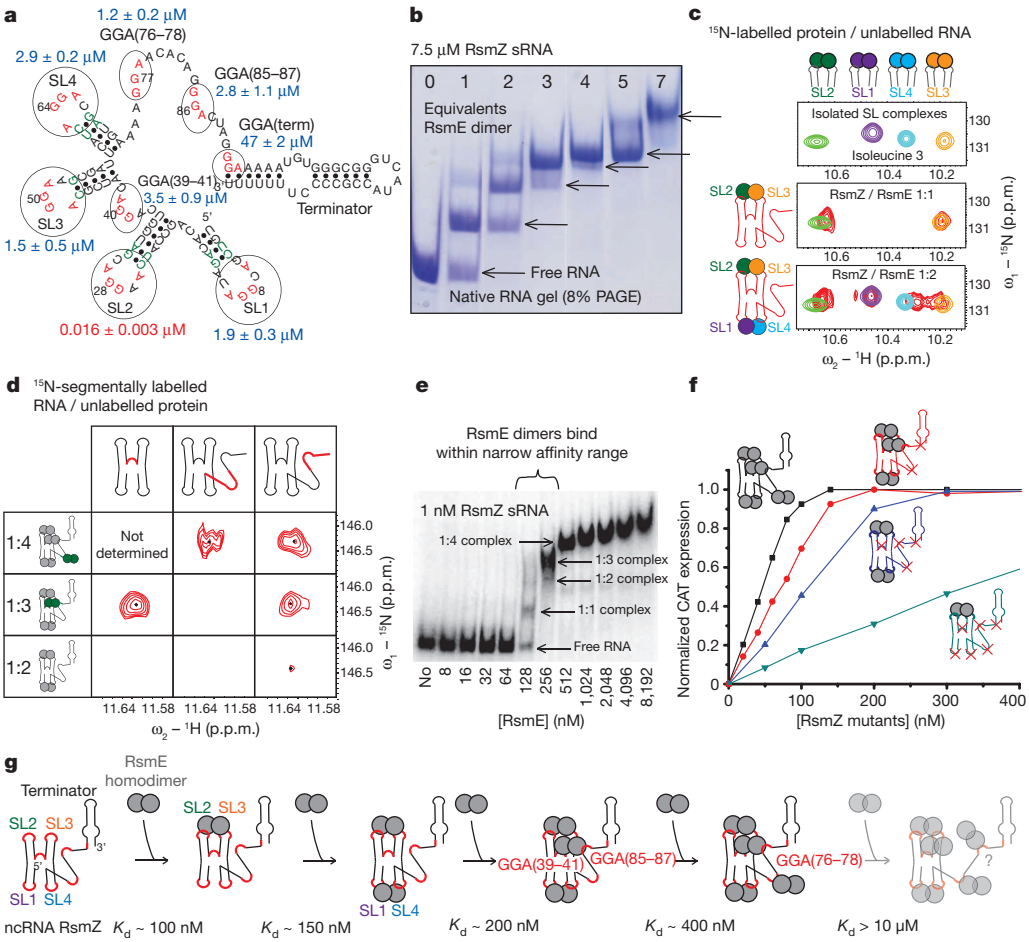


Figure 1 | Sequential and specific binding of the RsmE protein to the RsmZ sRNA. **a**, SL2 is the only high-affinity binding site in isolation. **b**, Up to five RsmE protein dimers can be accommodated by the RsmZ sRNA. **c**, **d**, Monitoring the binding of the first and second (**c**) and third and fourth (**d**) RsmE dimers by ^{1}H–^{15}N HSQC NMR spectroscopy observing the protein or RNA side, respectively. **e**, Gel shift assay. **f**, *In vitro* translation assay demonstrating the functional importance for the binding of the four dimers with a submicromolar K_d. **g**, Schematic representation of RsmZ/RsmE RNP assembly.

dissociation constants (Fig. 1a), this indicates that RsmE binding to RsmZ is cooperative. To determine which RsmE dimer binding sites on RsmZ are functionally important for activating translation initiation, we performed *in vitro* translation assays using a chloramphenicol acetyltransferase reporter gene fused to the 5′-untranslated region of the *hcnA* mRNA (Fig. 1f). Addition of RsmE at various concentrations of RsmZ sRNAs with mutations in different GGA motifs clearly demonstrated that the binding of all four dimers in the submicromolar affinity range are functionally important to reach optimal translation activation.

70-kDa RsmE–RsmZ solution structures

Although ordered, hierarchical and cooperative ribonucleoprotein (RNP) assembly is generally under evolutionary selection[13] as exemplified by the ribosome[14,15], the spliceosome[16] or telomerase[17] RNP assemblies, it is rather unexpected to find that a sRNA such as RsmZ appears to be bound cooperatively by several RsmE dimers in such a well-defined and sequential manner considering that the primary role of this sRNA should be simply to sequester many CsrA/RsmE proteins. Does such an ordered assembly also lead to a well-defined tertiary structure that could reveal the molecular basis of the cooperative binding? To this end, we determined the 70-kDa solution structure of RsmZ(1−72) in complex with three RsmE protein dimers (Fig. 2a and Supplementary Discussion) using a novel approach for structure determination of large protein–RNA complexes in solution that combines NMR and EPR spectroscopy (Extended Data Fig. 1 and Methods)[18].

We first determined the structures of the five RsmE–RNA subcomplexes (RsmE bound to the four stem–loops and to the linker between SL2 and SL3, each in isolation) at high resolution using NMR spectroscopy (Extended Data Fig. 1, step 1)[10]. Second, we checked experimentally that each individual subcomplex has the same structure in the full complex by comparing the NMR chemical shifts of the respective complexes (step 2). Third, we measured 21 distances by pulsed EPR spectroscopy between spin labels attached on the RNA of each subcomplex and used them as long-range distance constraints between the subcomplex structures (step 3). This latter step in the structure calculation was particularly difficult as we could see evidence for two conformations simultaneously present in solution, as evidenced by the appearance of two sets of NMR cross-peaks for the RNA resonances in the linker regions (Extended Data Fig. 2b, c and Supplementary Discussion) and by the presence of two well-separated peaks in some EPR distance distributions (Fig. 2b), or by broad EPR distributions that cannot arise from a single static structure (Extended Data Fig. 3b). Nevertheless, we could discriminate the distance constraints belonging to each of the two conformers and obtain well-defined structural ensembles for both conformers (Fig. 2a and Extended Data Table 1).

Both conformations adopt RNA folds that have not, to our knowledge, been reported before. Although four stem–loops are present, no coaxial stacking is observed between any pairs of them, unlike what is found in the structures of large RNAs or RNPs[19]. Rather, the RNPs seem to be arranged in a way that maximizes the distances between the three

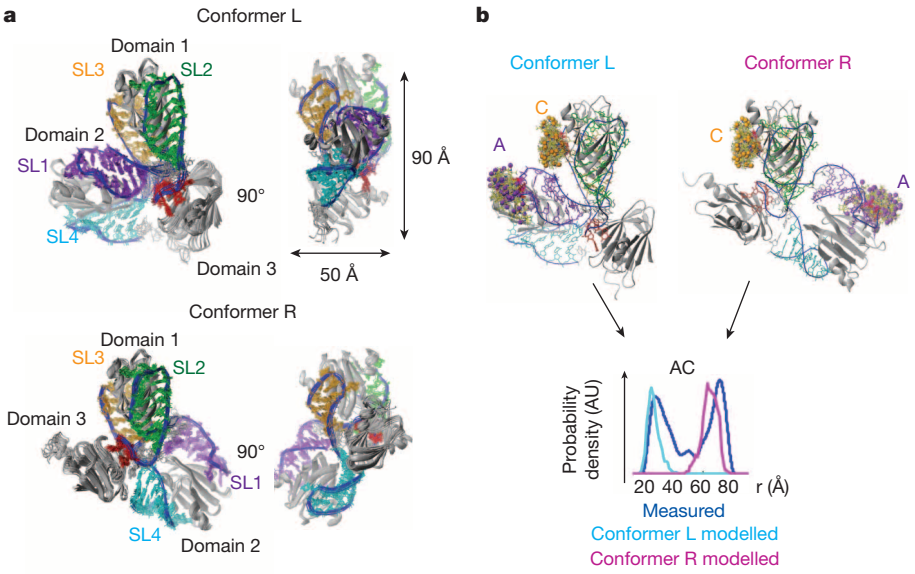


Figure 2 | Solution structures of both conformers of the 70-kDa complex between RsmZ(1−72) and three RsmE homodimers. **a**, Structural ensembles of both conformations present in solution. **b**, Representative EPR distance distribution indicating the presence of two conformations in solution. Shown are the modelled spin label conformations at the sites A and C in both conformers L and R.

bound RsmE dimers (Fig. 2a). The structure ensembles reveal tertiary interactions that stabilize their global fold such as intermolecular contacts between SL4 and the third RsmE dimer in conformer R or intramolecular interactions between the SL1–SL2 and SL2–SL3 linker residues (Extended Data Fig. 4). Both conformers of the RNP structure have the shape of an oblate disc with a diameter of around 90 Å and a thickness of 50 Å (Fig. 2a). Knowing the detailed assembly pathway helped to explain how this complex adopts two distinct and almost equally populated conformations, which differ in the relative position of domain 2 (SL1 and SL4) and domain 3 in respect to the first RsmE dimer (domain 1) (Fig. 2a). These two conformations cannot interconvert unless RsmE dissociates from SL2 and SL3, strongly suggesting two parallel assembly pathways and a conformational selection mechanism for the binding of the first RsmE dimer (Extended Data Fig. 4g). This means that the determined K_d values (Fig. 1e) are averaged over the two corresponding conformations (Supplementary Discussion). We can speculate that the presence of two parallel and productive sequestration pathways might provide conformational flexibility for the sRNA, allowing efficient sequestration of RsmE from the RBS in various topological mRNA contexts.

Positive and negative cooperativity

Almost the entire RsmZ sRNA becomes covered with proteins, illustrating the high binding capacity of this sRNA for the RsmE protein. The first three dimers all bind within a narrow affinity range (K_d of 100, 150 and 200 nM) despite very different individual binding sites. How is this tuned? The first RsmE protein dimer binds with a K_d of 100 nM, which might seem a lower affinity than expected considering that binding affinities of RsmE to the individual SL2 and SL3 were measured to be 16 nM and 1.5 μM K_d, respectively (Fig. 1). We can explain this unexpected result by two observations. First, when studying the binding of the high-affinity SL2 in isolation, we saw by isothermal titration calorimetry (ITC) and NMR titration experiments that binding of the first dimer was accompanied by a negative allosteric effect reducing the affinity for binding of the second dimer (Fig. 3a, Extended Data Fig. 5 and Supplementary Discussion). Second, the linker between SL2 and SL3 does not have the optimal length to reach the highest affinity (Fig. 3b). A sRNA mutant in which the SL2–SL3 linker was shortened by three nucleotides resulted in more than threefold higher affinity for RsmE and enhanced translation activation compared to the wild-type linker (Fig. 3b). Overall, an inherent property of the RsmE dimer and the non-optimal linker length between SL2 and SL3 explain how the binding of the first dimer is tuned to only 100 nM K_d. Our findings also suggest that the RNA linker between SL2 and SL3 in RsmZ has evolved to contain an additional binding site (domain 3) rather than to maximize the binding affinity and translation activation potential of the first dimer.

In contrary to the first RsmE dimer, the following two RsmE dimers bind with much higher affinity compared to the binding of isolated sites, indicating a positive cooperativity (Fig. 1). Notably, the binding affinity of the third dimer is strongly enhanced (25-fold, Extended Data Fig. 6a, b) by the binding of the first two dimers, which can be explained by two structural features. First, binding of the first RsmE dimer to SL2 and SL3 fixes the positions of the 5′ and 3′ ends of the third binding site and thus pre-form the latter (Fig. 3c). Second, in both conformers, binding of the second RsmE dimer to SL4 positions the 3′ end of the sRNA such that $GGA_{85–87}$ is brought into close proximity to the second binding site on the third RsmE dimer (Fig. 3d). Overall, our findings demonstrate that the sRNA RsmZ has an architecture permitting a specific and sequential protein binding in which the binding of an RsmE dimer pre-organizes the binding of a consecutive RsmE dimer. This precise RNA scaffold is conserved in *Pseudomonas* (Extended Data Fig. 7), supporting the functional importance for the cooperative binding of multiple RsmE proteins within a narrow affinity range of 100–200 nM K_d.

RNase protection and RsmE release

The presence of two compact and well-defined structures poses the question of how such a RNP could be turned over once translation activation is no longer required and the cell must return to its initial homeostasis phase. Bacterial RNA turnover is usually initiated by an endoribonucleolytic cleavage performed by RNase E[20,21]. In the free RNA, we could identify three specific cleavage sites that quite remarkably are located near GGA motifs (Fig. 4a and Extended Data Fig. 7e) and that are conserved in *Pseudomonas* (Extended Data Fig. 7d). We could further show that RsmE binding beyond the 1:1 stoichiometry progressively protects RsmZ from degradation by RNase E *in vitro* (Fig. 4b) and in *E. coli* whole-cell extracts containing the entire bacterial degradation machinery (Fig. 4c), consistent with previous *in vivo* findings[22]. In agreement with our RsmE binding data (Fig. 1), four RsmE dimers are required to strongly protect RsmZ from degradation (Fig. 4b–d). Our structural models explain the protection at this 4:1 stoichiometry as all three cleavage sites are bound by the protein (Fig. 4d). Thus, RsmZ binding by several RsmE dimers is not only an efficient way to sequester proteins but also to store them (Fig. 4c). Moreover, the RNA cleavage pattern proposes how RsmE could

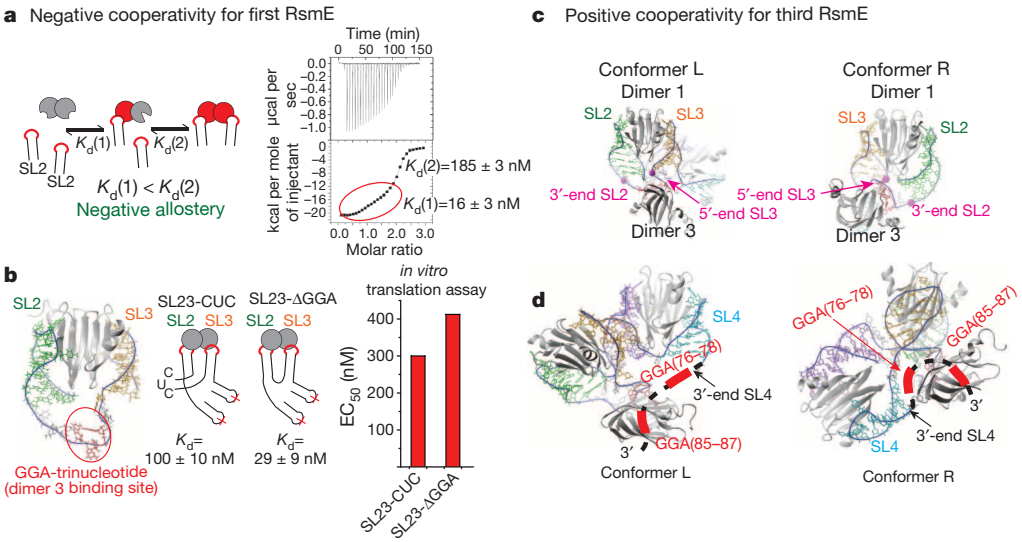


Figure 3 | Fine-tuning of the binding affinity of RsmZ for RsmE results from both negative and positive binding cooperativity. **a**, A negative allosteric effect is induced upon binding of a first high-affinity SL2 to the RsmE protein homodimer. **b**, The SL23 linker is longer than required to bind dimer 1 with the highest possible affinity (as measured by ITC; see Methods). EC_{50}, half-maximal effective concentration. **c**, Binding of the first dimer preforms the binding site of the third dimer. **d**, The 3′ end of SL4 is close to the third RsmE dimer, facilitating its binding to the $GGA_{85–87}$ binding motif.

be released from RsmZ. The increased transcription of RsmZ or other RsmE titrating sRNAs or mRNAs would result in dissociation of the RsmZ-bound RsmE and its redistribution onto the newly transcribed sRNAs or mRNAs (at constant RsmE protein concentration[22]). The concomitance of some of the sites for RsmE binding and RNase E cleavage suggests that an RsmE dissociation followed by RNase E cleavage will weaken RsmE re-association and therefore lead to an unidirectional RsmE release (Figs 4d and 5). Therefore, RsmZ can be seen as a true protein 'sponge' because this ncRNA can sequester, store and, coupled with RNase E, release proteins (Fig. 5).

Discussion

Here, we have structurally investigated how the ncRNA RsmZ sequesters the RsmE protein in order to activate translation. Unexpectedly, RsmZ binds sequentially up to five RsmE dimers, most of them cooperatively within a narrow affinity range (100–200 nM K_d). Other sRNAs of the Csr/Rsm family have entirely different secondary structures, resulting in various assembly pathways and three-dimensional structures. As a consequence, different sRNAs have distinct sequestration properties and therefore different physiological functions, illustrated by the finding that RsmZ, but not RsmY, overexpression coincides with impaired biofilm development in *P. aeruginosa*[23]. Thus, it is not surprising that many bacterial species contain different types of CsrA/RsmE sequestrating sRNAs, which are expressed at different phases during cell growth[24] and can act either additively, redundantly or independently on different mRNA targets[23,25,26], thereby causing different physiological reactions[23,25].

This evolutionary conserved sequential and ordered assembly has several advantages compared to a random sequestration process. It prevents the formation of long-lived misassembled intermediates or aggregates[13,17], which could either have poor sequestration capacity or be too stable for eventual degradation. Furthermore, this hierarchical assembly equips the sRNA with a well-defined and optimized binding process for RsmE that has been evolutionarily tuned to maximize its sequestration capacity and modulate the concentration range in which it efficiently sequesters RsmE. This sequential assembly also allows the storage of the protein, the RNP becoming insensitive to degradation

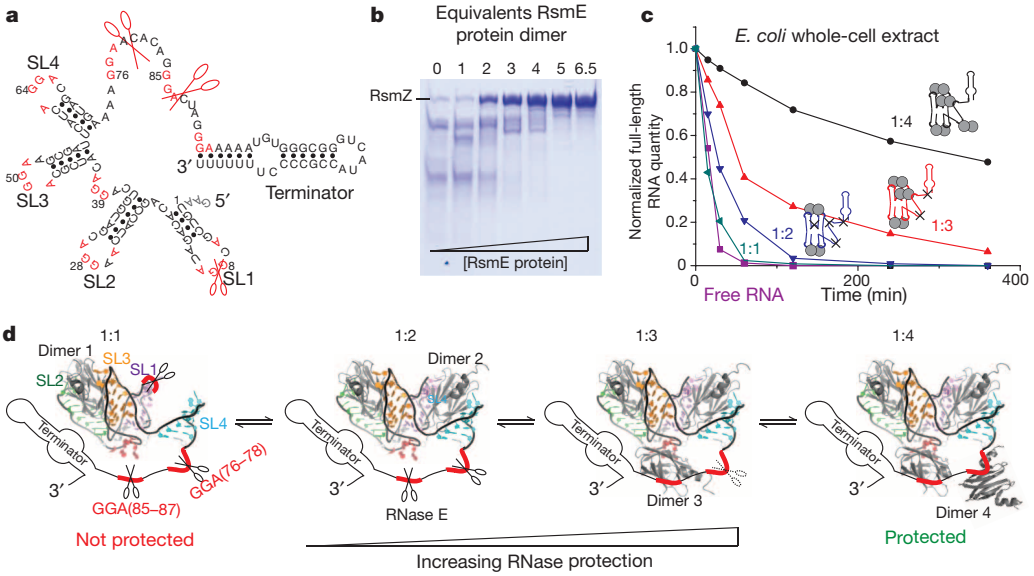


Figure 4 | RsmE progressively protects the sRNA RsmZ from degradation. **a**, Specific cleavage of RsmZ sRNA by RNase E. **b**, **c**, Protection of sRNA RsmZ by the RsmE protein against RNase E (**b**) or *E. coli* whole-cell extract (**c**). **d**, Structural basis for sRNA RsmZ protection (shown for conformer L).

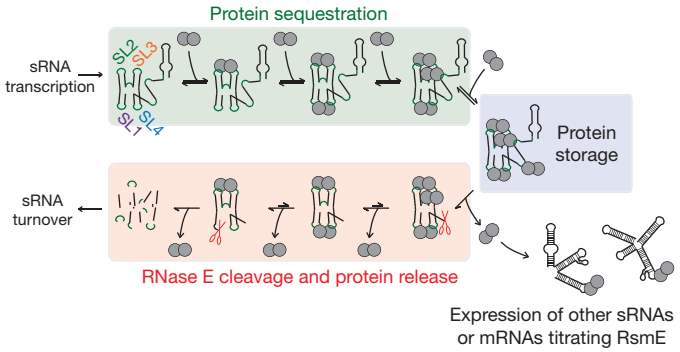


Figure 5 | The life of the ncRNA RsmZ as a protein sponge. The sRNA RsmZ is transcribed and then sequesters the RsmE protein. When bound to RsmE, the sRNA is stable and can therefore store the protein. The transcription of more RsmZ sRNA or other RsmE-titrating sRNAs (RsmX or RsmY) or mRNAs (*hcnA* mRNA) could result in RsmE protein release from the RsmE-bound RsmZ sRNA, permitting its subsequent degradation.

after a 4:1 ratio of protein to RNA. Finally, coupling between the binding sites for RsmE and RNase E in the sRNA enables the release of the sequestered RsmE proteins from the RNP. The capacity for an RNA to sequester, store and release RNA is reminiscent of the recently discovered circular RNAs that act as microRNA sponges[27,28]. By analogy, RsmZ can be seen in light of our findings as a protein sponge (Fig. 5).

The mode of RNP assembly described here for RsmE is radically different from other cooperative assemblies found to date involving the binding of identical proteins such as hnRNPA1 (ref. 29) or the HIV-1 Rev[30] to RNA. Cooperative binding of RsmE to RsmZ does not involve protein–protein interactions but rather allosteric changes in the RNA allowing subsequent protein binding, a mode of binding that is more reminiscent to bacterial ribosome assembly[14,15]. As a comparison, it will be interesting to see whether the circular RNA sponges sequester microRNA in a random fashion or also bind in a cooperative, well-defined and regulated manner.

METHODS SUMMARY

RsmE protein expression and purification[7,10], RNA transcription, purification and preparation of segmental isotopically labelled[11,12] or spin-labelled RNAs[18] were performed as previously described or described in Methods. ITC binding experiments were conducted as described in Methods. EPR, NMR spectroscopy and the structure calculation are detailed in Methods, Extended Data Figs 1–3, 8 and 9, Extended Data Table 1 and Supplementary Methods and are described in more detail elsewhere[10,18]. *In vitro* translation[31] and RNase protection assays are outlined in Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 October 2013; accepted 24 March 2014.

Published online 14 May 2014.

1. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136,** 629–641 (2009).
2. Storz, G., Vogel, J. & Wassarman, K. M. Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell* **43,** 880–891 (2011).
3. Waters, L. S. & Storz, G. Regulatory RNAs in bacteria. *Cell* **136,** 615–628 (2009).
4. Papenfort, K. & Vogel, J. Regulatory RNA in bacterial pathogens. *Cell Host Microbe* **8,** 116–127 (2010).
5. Romeo, T., Vakulskas, C. A. & Babitzke, P. Post-transcriptional regulation on a global scale: form and function of Csr/Rsm systems. *Environ. Microbiol.* **15,** 313–324 (2013).
6. Lapouge, K., Schubert, M., Allain, F. H. & Haas, D. Gac/Rsm signal transduction pathway of γ-proteobacteria: from RNA recognition to regulation of social behaviour. *Mol. Microbiol.* **67,** 241–253 (2008).
7. Schubert, M. *et al.* Molecular basis of messenger RNA recognition by the specific bacterial repressing clamp RsmA/CsrA. *Nature Struct. Mol. Biol.* **14,** 807–813 (2007).
8. Dubey, A. K., Baker, C. S., Romeo, T. & Babitzke, P. RNA sequence and secondary structure participate in high-affinity CsrA-RNA interaction. *RNA* **11,** 1579–1587 (2005).
9. Heeb, S., Blumer, C. & Haas, D. Regulatory RNA as mediator in GacA/RsmA-dependent global control of exoproduct formation in *Pseudomonas fluorescens* CHA0. *J. Bacteriol.* **184,** 1046–1056 (2002).
10. Duss, O., Michel, E., Diarra Dit Konte, N., Schubert, M. & Allain, F. H. Molecular basis for the wide range of affinity found in Csr/Rsm protein–RNA recognition. *Nucleic Acids Res.* http://dx.doi.org/10.1093/nar/gku141 (21 February 2014).
11. Duss, O., Lukavsky, P. J. & Allain, F. H. Isotope labeling and segmental labeling of larger RNAs for NMR structural studies. *Adv. Exp. Med. Biol.* **992,** 121–144 (2012).
12. Duss, O., Maris, C., von Schroetter, C. & Allain, F. H. A fast, efficient and sequence-independent method for flexible multiple segmental isotope labeling of RNA using ribozyme and RNase H cleavage. *Nucleic Acids Res.* **38,** e188 (2010).
13. Marsh, J. A. *et al.* Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* **153,** 461–470 (2013).
14. Woodson, S. A. RNA folding pathways and the self-assembly of ribosomes. *Acc. Chem. Res.* **44,** 1312–1319 (2011).
15. Shajani, Z., Sykes, M. T. & Williamson, J. R. Assembly of bacterial ribosomes. *Annu. Rev. Biochem.* **80,** 501–526 (2011).
16. Wahl, M. C., Will, C. L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136,** 701–718 (2009).
17. Stone, M. D. *et al.* Stepwise protein-mediated RNA folding directs assembly of telomerase ribonucleoprotein. *Nature* **446,** 458–461 (2007).
18. Duss, O., Yulikov, M., Jeschke, G. & Allain, F. H. EPR-aided approach for solution structure determination of large RNAs or protein–RNA complexes. *Nature Commun.* http://dx.doi.org/10.1038/ncomms4669 (2014).
19. Holbrook, S. R. Structural principles from large RNAs. *Annu. Rev. Biophys.* **37,** 445–464 (2008).
20. Mackie, G. A. RNase E: at the interface of bacterial RNA processing and decay. *Nature Rev. Microbiol.* **11,** 45–57 (2013).
21. Carpousis, A. J., Luisi, B. F. & McDowall, K. J. Endonucleolytic initiation of mRNA decay in *Escherichia coli*. *Prog. Mol. Biol. Transl. Sci.* **85,** 91–135 (2009).
22. Reimmann, C., Valverde, C., Kay, E. & Haas, D. Posttranscriptional repression of GacS/GacA-controlled genes by the RNA-binding protein RsmE acting together with RsmA in the biocontrol strain *Pseudomonas fluorescens* CHA0. *J. Bacteriol.* **187,** 276–285 (2005).
23. Petrova, O. E. & Sauer, K. The novel two-component regulatory system BfiSR regulates biofilm development by controlling the small RNA *rsmZ* through CafA. *J. Bacteriol.* **192,** 5275–5288 (2010).
24. Kay, E. *et al.* Two GacA-dependent small RNAs modulate the quorum-sensing response in *Pseudomonas aeruginosa*. *J. Bacteriol.* **188,** 6026–6033 (2006).
25. Bordi, C. *et al.* Regulatory RNAs and the HptB/RetS signalling pathways fine-tune *Pseudomonas aeruginosa* pathogenesis. *Mol. Microbiol.* **76,** 1427–1443 (2010).
26. Kay, E., Dubuis, C. & Haas, D. Three small RNAs jointly ensure secondary metabolism and biocontrol in *Pseudomonas fluorescens* CHA0. *Proc. Natl Acad. Sci. USA* **102,** 17136–17141 (2005).
27. Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* **495,** 384–388 (2013).
28. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495,** 333–338 (2013).
29. Okunola, H. L. & Krainer, A. R. Cooperative-binding and splicing-repressive properties of hnRNP A1. *Mol. Cell. Biol.* **29,** 5620–5631 (2009).
30. Daugherty, M. D., Liu, B. & Frankel, A. D. Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. *Nature Struct. Mol. Biol.* **17,** 1337–1342 (2010).
31. Michel, E. & Wüthrich, K. High-yield *Escherichia coli*-based cell-free expression of human proteins. *J. Biomol. NMR* **53,** 43–51 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank G. A. Mackie for providing the RNase E (1−529) clone; N. Diarra Dit Konté and B. Alila for the help with RNA production; M. Blatter for help with structure calculations; G. Wider, T. Stahel, C. Maris and F. Damberger for help with NMR spectroscopy setup; G. Braach, Y. Nikolaev and M. Sattler for discussions. This work was supported by the Swiss National Science Foundation (SNF) grant numbers 3100A0-118118, 31003ab-133134 and 31003A-149921 to F.H.-T.A. and the SNF-NCCR structural biology Iso-lab.

Author Contributions O.D. prepared protein and RNA samples for NMR and EPR spectroscopy, collected and analysed the NMR experiments, designed the structure calculation protocol and performed structure calculations, conducted and analysed ITC experiments, gel shift and RNase E cleavage assays, designed the project and the experiments and wrote the manuscript. E.M. designed, performed and analysed *in vitro* translation and *E. coli* whole-cell extract RNase protection assays. M.Y. measured and analysed EPR data. M.S. performed initial NMR experiments. G.J. analysed EPR data and secured funding. F.H.-T.A. designed and supervised the study, secured funding, and wrote the manuscript. All authors discussed the results, commented on and approved the manuscript.

Author Information Coordinates, NMR and EPR restraints have been deposited in the Protein Data Bank under accession numbers 2mf0 and 2mf1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.H.-T.A. (allain@mol.biol.ethz.ch).

LETTER

doi:10.1038/nature13254

Three regimes of extrasolar planet radius inferred from host star metallicities

Lars A. Buchhave[1,2], Martin Bizzarro[2], David W. Latham[1], Dimitar Sasselov[1], William D. Cochran[3], Michael Endl[3], Howard Isaacson[4], Diana Juncher[2,5] & Geoffrey W. Marcy[4]

Approximately half of the extrasolar planets (exoplanets) with radii less than four Earth radii are in orbits with short periods[1]. Despite their sheer abundance, the compositions of such planets are largely unknown. The available evidence suggests that they range in composition from small, high-density rocky planets to low-density planets consisting of rocky cores surrounded by thick hydrogen and helium gas envelopes. Here we report the metallicities (that is, the abundances of elements heavier than hydrogen and helium) of more than 400 stars hosting 600 exoplanet candidates, and find that the exoplanets can be categorized into three populations defined by statistically distinct ($\sim 4.5\sigma$) metallicity regions. We interpret these regions as reflecting the formation regimes of terrestrial-like planets (radii less than 1.7 Earth radii), gas dwarf planets with rocky cores and hydrogen–helium envelopes (radii between 1.7 and 3.9 Earth radii) and ice or gas giant planets (radii greater than 3.9 Earth radii). These transitions correspond well with those inferred from dynamical mass estimates[2,3], implying that host star metallicity, which is a proxy for the initial solids inventory of the protoplanetary disk, is a key ingredient regulating the structure of planetary systems.

Shortly after the discovery of the first exoplanets, host star metallicity was suggested to have a role in the formation of planetary systems[4]. Indeed, the well-established tendency for hot Jupiters to be more frequently found orbiting metal-rich stars has been confirmed by a number of studies[5,6]. Although it has recently been shown that small planets form for a wide range of host star metallicities[7–11], it is clear that the metallicities of stars with small planets are on average lower than those of gas giants. This suggests that subtle differences may exist in the metallicities of the host stars of small exoplanets, and this, in turn, may be linked to distinct physical properties of the underlying planet populations. However, effectively probing this regime requires a large sample of homogeneously derived metallicities for stars with small planets. Therefore, using our stellar parameters classification (SPC) tool[7] (Methods Summary), we analyse more than 2,000 high-resolution spectra of Kepler Objects of Interest[12] (KOIs) gathered by the Kepler Follow-up Program, yielding precise stellar parameters (provided as a table in machine-readable form), including metallicities, of 405 stars orbited by 600 exoplanet candidates.

Our sample of spectroscopic metallicities of stars hosting small planets is a factor of two larger than any previous sample[7], allowing us to probe in greater detail for significant differences in the metallicities of stars hosting planets of different sizes. At various radii, we divide the sample into two bins of stars hosting small and large planets, and perform a two-sample Kolmogorov–Smirnov test to determine whether the metallicities of the two distributions of host stars are not drawn randomly from the same parent population. We find two significant features in the Kolmogorov–Smirnov test diagram, one at $1.7R_\oplus$ (Earth radii) with a significance of 4.5σ and one at $3.9R_\oplus$ with a significance of 4.6σ, suggesting transitions between three exoplanet size regimes (Fig. 1). The average metallicity of the host stars increases with planet size, yielding average metallicities of -0.02 ± 0.02, 0.05 ± 0.01 and 0.18 ± 0.02 dex in the respective regimes. To assess the uncertainty in radius at which these transitions occur, we perform a Monte Carlo simulation by drawing 10^6 sets of data, where the host star metallicities and planetary radii are randomly perturbed within the uncertainties (the uncertainty in the planetary radius is assumed to be dominated by the uncertainty in the radius of the host star). We find the two features to be at $1.55^{+0.88}_{-0.04}R_\oplus$ ($4.2^{+0.5}_{-0.4}\sigma$) and $3.52^{+0.74}_{-0.28}R_\oplus$ ($4.7^{+0.6}_{-0.4}\sigma$), consistent with the original data.

Small planets with short periods could undergo significant evaporation of their atmospheres[13]. These planets will therefore not obey the radius–metallicity relation we are studying, because any accumulated gas would have evaporated. Therefore, we remove small ($R_P < 3R_\oplus$, where R_P is the exoplanet radius), highly irradiated planets (stellar flux, $F_\nu > 5 \times 10^5\,\mathrm{J\,s^{-1}\,m^{-2}}$) from the sample, leaving 463 planets orbiting 324 stars, which increases the significance of the feature at $1.7R_\oplus$ from 3.5σ to the reported 4.5σ. For comparison, the rocky planet Kepler-10b, with an 0.8-d period[14], receives a flux of $F_\nu \approx 48 \times 10^5\,\mathrm{J\,s^{-1}\,m^{-2}}$, whereas Kepler-11c[15], whose density suggests it is gaseous, receives $F_\nu \approx 1.3 \times 10^5\,\mathrm{J\,s^{-1}\,m^{-2}}$.

Recent studies suggest that the masses and radii of small planets ($1.5R_\oplus$–$4R_\oplus$) follow a linear relationship, implying that planet density decreases with increasing planet radius[2]. However, this relationship must change significantly for larger planets ($>4R_\oplus$) to explain the large mass of gas giant planets such as Jupiter[2]. Our data indicate a statistically significant increase in metallicity at a comparable planetary radius, $R_P = 3.9R_\oplus$. This observation is in agreement with the well-established correlation between a star's metallicity and its likelihood to host hot Jupiters[5,6], confirming that the formation regime for larger planets ($>3.9R_\oplus$) requires exceptionally high metallicity environments[7]. We therefore interpret the regime of larger planets ($R_P > 3.9R_\oplus$) to consist of ice and gas giant planets formed beyond the 'snow line' at around $\sim$3 AU (1 AU is the average Sun–Earth distance), where the availability of solids is a factor of four higher because volatile elements are able to condense and form solids at cooler temperatures[16]. The high concentration of heavy elements in the protoplanetary disk, resulting from the higher metallicity and the condensation of volatiles, and the planet's large distance to the host star allow these planets to grow rapidly and amass a gaseous atmosphere before the gas in the protoplanetary disk dissipates. These planets then migrate to their present positions much closer to their host stars[17], yielding the hot Jupiters seen orbiting stars with high metallicities.

To explore the implications of the feature at $1.7R_\oplus$ in the metallicity–radius plane (Fig. 1), we compare our results with recent work attempting to determine the radius at which the transition from gaseous to rocky planets occurs. Dynamical masses derived from precise radial velocities of transiting exoplanets indicate that planets with $R_P > 2R_\oplus$ have densities that imply increasing amounts by volume of light material, whereas planets with $R_P < 1.5R_\oplus$ have densities systematically

[1]Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA. [2]Centre for Star and Planet Formation, Natural History Museum of Denmark, University of Copenhagen, DK-1350 Copenhagen, Denmark. [3]McDonald Observatory, The University of Texas, Austin, Texas 78712, USA. [4]University of California, Berkeley, California 94720, USA. [5]Niels Bohr Institute, University of Copenhagen, DK-2100 Copenhagen, Denmark.

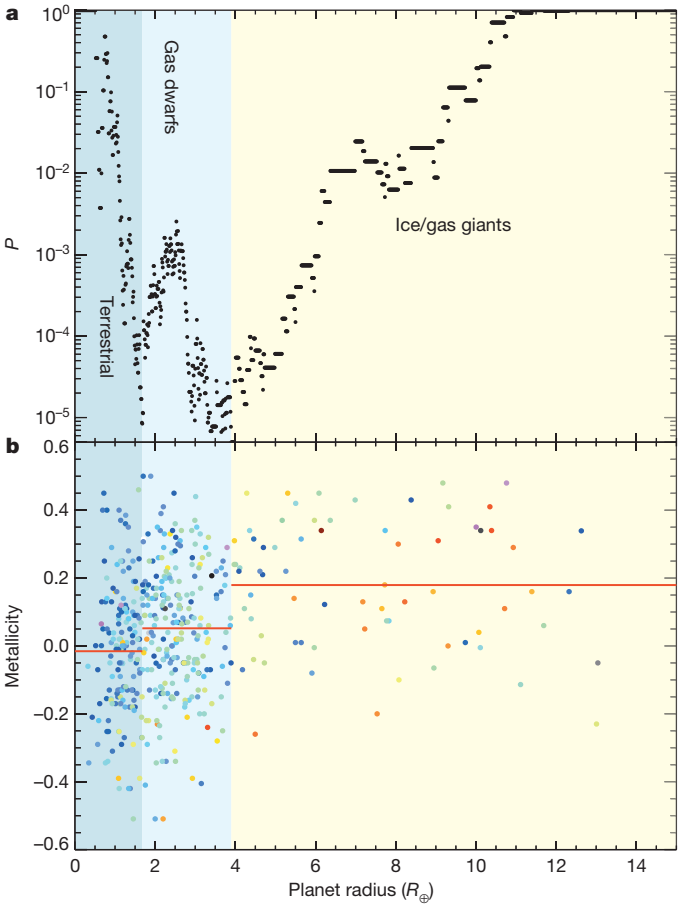


Figure 1 | Host star metallicities and three types of exoplanets with different composition. **a**, *P* value of the two-sample Kolmogorov–Smirnov test. **b**, Radii of the individual planets and their host star metallicities. Point colour represents the logarithm of the period of the planets (blue, shortest period; red, longest period). The solid red lines are the average metallicities in the three regions (-0.02 ± 0.02, 0.05 ± 0.01 and 0.18 ± 0.02 dex, where each uncertainty is 1 s.e.m. of the host star metallicities in the corresponding bin).

greater than that of Earth[2]. Moreover, an analysis of data for a larger sample of planets (including masses derived from transit timing variations), has shown that planets with $R_P < 1.5R_\oplus$ probably are of rocky composition[3]. Finally, it has been suggested that $R_P \approx 1.75R_\oplus$ is a physically motivated transition point between rocky and gaseous planets, based on reported masses and radii combined with thermal evolutionary atmosphere models[18]. The statistically significant peak in the metallicity–radius plane at $1.7R_\oplus$ agrees with these findings, suggesting that the compositions of small exoplanets ($R_P < 3.9R_\oplus$) in close proximity to their host stars are also regulated by the number density of solids in the protoplanetary disk. Thus, we interpret the two regimes of smaller planets identified by the host star metallicities as reflecting the transition between rocky terrestrial exoplanets that have not amassed a gaseous atmosphere ($R_P < 1.7R_\oplus$) and planets with rocky cores that have accumulated an envelope of hydrogen, helium and other volatiles, which we denote gas dwarfs ($1.7R_\oplus < R_P < 3.9R_\oplus$).

The formation mechanism of the terrestrial and gas dwarf exoplanet regimes in short orbital periods is not fully understood. In one model, these small exoplanets are believed to form *in situ* with little post-assembly migration[19,20]. Although the *in situ* accretion model seems to be successful in reproducing the observed distribution of the 'hot Neptune' and super-Earth systems, including their orbital spacing[21], it requires unusually large amounts of solids in the innermost protoplanetary disk. A competing model invokes accretion during the inward migration of a population of planetary embryos formed at a range of orbital distances beyond the snow line[22,23]. On this view, Mars- and Earth-size embryos migrate inwards owing to tidal interaction with the disk[24], and accumulate at the inner edge of the protoplanetary disk, where they complete their assembly[25]. Irrespective of the formation mechanism, however, the observed peak in the metallicity–radius plane at $1.7R_\oplus$ suggests that the final mass and composition of a small exoplanet is controlled by the amount of solid material available in the protoplanetary disk. A higher-metallicity environment promotes a more rapid and effective accretion process, thereby allowing the cores to amass a gaseous envelope before dissipation of the gas. In contrast, lower-metallicity environments may result in the assembly of rocky cores of several Earth masses on timescales greater than that inferred for gas dispersal in protoplanetary disks[19] (<10 Myr), yielding cores without gaseous hydrogen–helium atmospheres.

A prediction from gas accretion on short orbital periods is that the critical mass at which a core can accrete an atmosphere is $M_{cr} \approx 2.6M_\oplus(\eta/0.3)^{1/2}(P_{orb}/1\ d)^{5/12}$, where $\eta = M_{atm}/M_{cr}$ is the fractional mass comprised by the atmosphere[26]. In this model, the planetary mass and, thus, radius indicating the transition from rocky to gaseous planets should increase with orbital period. However, the exact opposite dependence, namely a decrease in core mass with increased orbital period, has also been suggested[27]. To investigate whether the radius of transition from rocky to gaseous planets found in our data shows a dependence on orbital period, we segregate the sample by period into four bins with approximately equal numbers of planets. We repeat the previously described Kolmogorov–Smirnov test for the planets in each of the period bins: we remove the larger planets ($R_P > 3.9R_\oplus$) and perform a Monte Carlo simulation by drawing 10^6 sets of data, where the host star metallicities and planetary radii are randomly perturbed within the uncertainties. The red line in Fig. 2 is a power-law fit to the transition radius, R_{cr}, inferred from our data ($R_{cr} = 1.06R_\oplus(P_{orb}/1\ day)^{0.17}$). Although additional data are required to confirm this relationship, the fit is apparently consistent with a critical core mass that increases with orbital period and an atmospheric fraction of 5% (ref. 26; blue dashed line in Fig. 2). If correct, this predicts the existence of more massive rocky exoplanets at longer orbital periods.

Although our analysis of a statistically significant number of planets and their host star metallicities allows us to distinguish between three distinct exoplanet regimes, we emphasize that a multitude of factors can affect the outcome of planet formation. Therefore, the transition radii inferred from our analysis probably represent gradual transitions

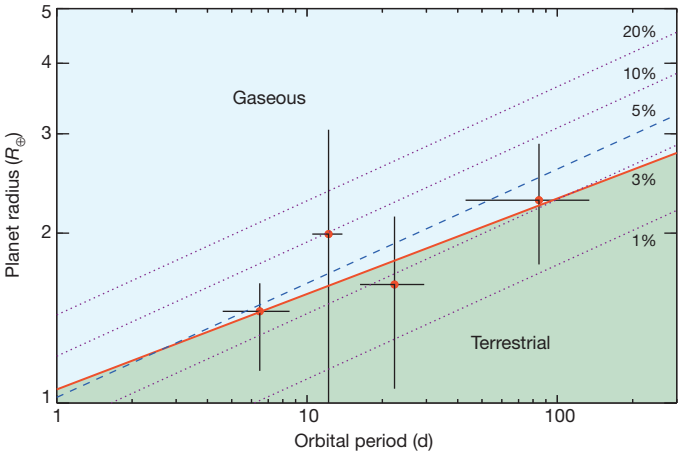


Figure 2 | The radius of transition from rocky to gaseous exoplanets. The transition radii (red points) are the means of the posterior distributions resulting from the Monte Carlo analysis (main text), and the error bars indicate the 1σ uncertainties. Using the mass–radius approximation[29] $M/M_\oplus = (R/R_\oplus)^{2.06}$, we plot the radius corresponding to M_{cr} with an atmosphere fraction of 5% as the blue dashed line and those for respective atmosphere fractions of 1%, 3%, 10% and 20% as the dotted purple lines. The solid red line is a power-law fit to the Monte Carlo data: $R_{cr} = 1.06R_\oplus(P_{orb}/1\ day)^{0.17}$.

between the different planet regimes and so may not apply to all planetary systems. However, the agreement between the transition radii inferred here and those deduced from dynamical mass measurements of transiting planets[2,3] implies that host star metallicity—and, by extension, the solids inventory of a protoplanetary disk—is one of the driving factors determining the outcome of planet formation.

METHODS SUMMARY

We use SPC[7] for the spectroscopic analysis yielding the stellar parameters in this work. SPC uses a grid of synthetic library spectra to derive the effective temperature, surface gravity, metallicity and rotational velocity simultaneously by matching the models with the observed spectra originating from a number of different instruments (Methods). The stellar parameters from SPC and Yonsei–Yale stellar evolutionary models[28] are used to estimate the radii of the host stars, which we couple with the photometric data from the Kepler mission to infer the planet radii. The majority of planets in the sample have short orbital periods, owing to the observational bias of Kepler towards shorter period planets. The mean and median periods are 38.0 and 12.4 d, respectively.

To investigate whether the radius of transition from rocky to gaseous planets found in our data shows a dependence on orbital period (Fig. 2), we remove the larger planets ($R_P > 3.9R_\oplus$) and segregate the remainder of the sample by period into four bins with approximately equal numbers of planets. The previously described Kolmogorov–Smirnov test (Fig. 1) is repeated for each bin by performing a Monte Carlo simulation in which the host star metallicities and planetary radii are randomly perturbed within the uncertainties. The data points plotted in Fig. 2 are the means of the resulting posterior distributions, and the error bars indicate the 1σ uncertainties.

To ensure that the statistical analyses are sound, we perform further tests evaluating the effects of uneven sampling and contamination (Methods). We find results consistent with those presented in the paper within the uncertainties, and conclude that our statistical analyses are robust.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 November 2013; accepted 12 March 2014.

1. Fressin, F. *et al.* The false positive rate of Kepler and the occurrence of planets. *Astrophys. J.* **766,** 81 (2013).
2. Marcy, G. W. *et al.* Masses, radii, and orbits of small Kepler planets: the transition from gaseous to rocky planets. *Astrophys. J. Suppl. Ser.* **210,** 20 (2014).
3. Weiss, L. M. & Marcy, G. W. The mass-radius relation for 65 exoplanets smaller than 4 Earth radii. *Astrophys. J.* **783,** L6 (2014).
4. Gonzalez, G. The stellar metallicity-giant planet connection. *Mon. Not. R. Astron. Soc.* **285,** 403–412 (1997).
5. Santos, N. C., Israelian, G. & Mayor, M. Chemical analysis of 8 recently discovered extra-solar planet host stars. *Astron. Astrophys.* **363,** 228–238 (2000).
6. Fischer, D. A. & Valenti, J. The planet-metallicity correlation. *Astrophys. J.* **622,** 1102–1117 (2005).
7. Buchhave, L. A. *et al.* An abundance of small exoplanets around stars with a wide range of metallicities. *Nature* **486,** 375–377 (2012).
8. Sousa, S. G., Santos, N. C., Israelian, G., Mayor, M. & Udry, S. Spectroscopic stellar parameters for 582 FGK stars in the HARPS volume-limited sample. Revising the metallicity-planet correlation. *Astron. Astrophys.* **533,** A141 (2011).
9. Sousa, S. G. *et al.* Spectroscopic parameters for 451 stars in the HARPS GTO planet search program. Stellar [Fe/H] and the frequency of exo-Neptunes. *Astron. Astrophys.* **487,** 373–381 (2008).
10. Ghezzi, L. *et al.* Stellar parameters and metallicities of stars hosting Jovian and Neptunian mass planets: a possible dependence of planetary mass on metallicity. *Astrophys. J.* **720,** 1290–1302 (2010).
11. Everett, M. E., Howell, S. B., Silva, D. R. & Szkody, P. Spectroscopy of faint Kepler mission exoplanet candidate host stars. *Astrophys. J.* **771,** 107 (2013).
12. Batalha, N. M. *et al.* Planetary candidates observed by Kepler. III. Analysis of the first 16 months of data. *Astrophys. J. Suppl. Ser.* **204,** 24 (2013).
13. Owen, J. E. & Wu, Y. Kepler planets: a tale of evaporation. *Astrophys. J.* **775,** 105 (2013).
14. Batalha, N. M. *et al.* Kepler's first rocky planet: Kepler-10b. *Astrophys. J.* **729,** 27 (2011).
15. Lissauer, J. J. *et al.* All six planets known to orbit Kepler-11 have low densities. *Astrophys. J.* **770,** 131 (2013).
16. Ida, S. & Lin, D. N. C. Toward a deterministic model of planetary formation. II. The formation and retention of gas giant planets around stars with a range of metallicities. *Astrophys. J.* **616,** 567–572 (2004).
17. Lin, D. N. C., Bodenheimer, P. & Richardson, D. C. Orbital migration of the planetary companion of 51 Pegasi to its present location. *Nature* **380,** 606–607 (1996).
18. Lopez, E. D. & Fortney, J. J. Understanding the mass-radius relation for sub-Neptunes: radius as a proxy for composition. Preprint at http://arxiv.org/abs/1311.0329 (2013).
19. Hansen, B. M. S. & Murray, N. Migration then assembly: formation of Neptune-mass planets inside 1 AU. *Astrophys. J.* **751,** 158 (2012).
20. Chiang, E. & Laughlin, G. The minimum-mass extrasolar nebula: in situ formation of close-in super-Earths. *Mon. Not. R. Astron. Soc.* **431,** 3444–3455 (2013).
21. Hansen, B. M. S. & Murray, N. Testing in situ assembly with the Kepler planet candidate sample. *Astrophys. J.* **775,** 53 (2013).
22. Terquem, C. & Papaloizou, J. C. B. Migration and the formation of systems of hot super-Earths and Neptunes. *Astrophys. J.* **654,** 1110–1120 (2007).
23. McNeil, D. S. & Nelson, R. P. On the formation of hot Neptunes and super-Earths. *Mon. Not. R. Astron. Soc.* **401,** 1691–1708 (2010).
24. Goldreich, P. & Tremaine, S. The excitation of density waves at the Lindblad and corotation resonances by an external potential. *Astrophys. J.* **233,** 857–871 (1979).
25. Raymond, S. N. & Cossou, C. No universal minimum-mass extrasolar nebula: Evidence against in-situ accretion of systems of hot super-Earths. Preprint at http://arxiv.org/abs/1401.3743 (2014).
26. Rafikov, R. R. Atmospheres of protoplanetary cores: critical mass for nucleated instability. *Astrophys. J.* **648,** 666–682 (2006).
27. Papaloizou, J. C. B. & Terquem, C. Critical protoplanetary core masses in protoplanetary disks and the formation of short-period giant planets. *Astrophys. J.* **521,** 823–838 (1999).
28. Yi, S. *et al.* Toward better age estimates for stellar populations: the Y^2 isochrones for solar mixture. *Astrophys. J. Suppl. Ser.* **136,** 417–437 (2001).
29. Lissauer, J. J. *et al.* Architecture and dynamics of Kepler's candidate multiple transiting planet systems. *Astrophys. J. Suppl. Ser.* **197,** 8 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements L.A.B. acknowledges support from the Harvard Origins of Life Initiative. M.B. acknowledges funding from the Danish National Research Foundation (grant number DNRF97) and from the European Research Council under ERC Consolidator grant agreement 616027- STARDUST2ASTEROIDS. D.W.L. acknowledges support from the Kepler Mission under NASA Cooperative Agreements NCC2-1390, NNX11AB99A and NNX13AB58A with the Smithsonian Astrophysical Observatory, and thanks the observers who helped obtain the TRES observations reported here, especially R. Stefanik, G. Esquerdo, P. Berlind and M. Calkins.

Author Contributions L.A.B. led the project and developed the classification tools for the metallicity analysis. M.B., D.W.L. and D.S. contributed to the discussion of the theoretical implications of the data. L.A.B., D.W.L., W.D.C., M.E., H.I., D.J. and G.W.M. worked on gathering the spectroscopic observations. All authors discussed the results and commented on the manuscript. L.A.B. and M.B. wrote the paper with input from D.W.L. and D.S.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.A.B. (lbuchhave@cfa.harvard.edu).

METHODS

Observations and stellar parameters. This study is based on stellar classifications by SPC[7] of 2,297 spectra observed using the Fibre-fed Echelle Spectrograph on the 2.6-m Nordic Optical Telescope on La Palma, Spain (488 spectra), the fibre-fed Tillinghast Reflector Echelle Spectrograph on the 1.5-m Tillinghast Reflector at the Fred Lawrence Whipple Observatory on Mt Hopkins, Arizona (985 spectra), the Tull Coudé Spectrograph on the 2.7-m Harlan J. Smith Telescope at the McDonald Observatory Texas (653 spectra) and the HIRES spectrograph on the 10-m Keck I telescope at Mauna Kea, Hawaii (171 spectra). We included only the most secure stellar classifications by limiting our sample to stars with effective temperatures of $4{,}800\,\mathrm{K} < T_{\mathrm{eff}} < 6{,}500\,\mathrm{K}$, projected rotational velocities of $v\sin(i) < 20\,\mathrm{km\,s^{-1}}$, spectra with signal-to-noise ratios per resolution element of more than 25, and normalized cross-correlation function peak heights of more than 0.9 (indicating the quality of the stellar classification).

We improve on the determination of the surface gravity, known to be prone to degeneracies with effective temperature and metallicity[30], by imposing a prior on the surface gravity from stellar evolutionary models and an initial estimate of the star's effective temperature and metallicity. This is particularly useful for cooler stars, where the evolutionary models put tight constraints on the surface gravity. We use the stellar parameters from SPC and the Yonsei–Yale stellar evolutionary models[28] to estimate the radii of the host stars, and, using the photometrically derived planet radii from Kepler, we correct the planetary radii based on the Kepler Input Catalogue photometry, which are known to be prone to systematic biases and large uncertainties. The improved stellar radii reduce the uncertainties in the planetary radii from an average error of 34% to one of 11%, assuming that the major contribution to the uncertainty in the planetary radii originates from the stellar radii.

Uneven sampling. To investigate the effect of uneven sample size, we performed a Monte Carlo test with 10^6 realizations where we randomly drew observations from the smaller of the two samples, making each of the two samples equal in size at all times. We find the ice or gas giant transition to be at $3.58^{+0.75}_{-0.38}R_{\oplus}$ with a significance of $4.9^{+0.5}_{-0.4}\sigma$ and the rocky transition to be at $1.60^{+0.83}_{-0.10}R_{\oplus}$ with a significance of $4.4^{+0.5}_{-0.4}\sigma$. Both values are consistent with the Monte Carlo analysis reported in the paper. We conclude that the uneven sample size does not affect the significance of our statistical analysis.

Contamination. We are using the two-sample Kolmogorov–Smirnov test for the statistical analysis, but we find three distinct populations of exoplanets. To establish whether contamination from the third sample affects our results, we remove the planets from the third sample (removing $R_{\mathrm{P}} > 3.9R_{\oplus}$ when searching for the peak at $1.7R_{\oplus}$ and removing $R_{\mathrm{P}} < 1.7R_{\oplus}$ for the peak at $3.9R_{\oplus}$) and subsequently carry out the Kolmogorov–Smirnov test in an attempt to recover each of the peaks in Fig. 1. We find the first transition to be at the same radius ($1.7R_{\oplus}$), albeit with a lower significance (3.1σ). We find the ice/gas giant transition close to the one reported in the paper, again with a slightly lower significance ($4.2R_{\oplus}$ at 3.9σ). Again, evaluation of our data using a different approach supports our results and conclusions.

30. Torres, G. *et al.* Improved spectroscopic parameters for transiting planet hosts. *Astrophys. J.* **757,** 161 (2012).

LETTER

doi:10.1038/nature13388

Direct high-precision measurement of the magnetic moment of the proton

A. Mooser[1,2]†, S. Ulmer[3], K. Blaum[4], K. Franke[3,4], H. Kracke[1,2], C. Leiteritz[1], W. Quint[5,6], C. C. Rodegheri[1,4], C. Smorra[3] & J. Walz[1,2]

One of the fundamental properties of the proton is its magnetic moment, μ_p. So far μ_p has been measured only indirectly, by analysing the spectrum of an atomic hydrogen maser in a magnetic field[1]. Here we report the direct high-precision measurement of the magnetic moment of a single proton using the double Penning-trap technique[2]. We drive proton-spin quantum jumps by a magnetic radio-frequency field in a Penning trap with a homogeneous magnetic field. The induced spin transitions are detected in a second trap with a strong superimposed magnetic inhomogeneity[3]. This enables the measurement of the spin-flip probability as a function of the drive frequency. In each measurement the proton's cyclotron frequency is used to determine the magnetic field of the trap. From the normalized resonance curve, we extract the particle's magnetic moment in terms of the nuclear magneton: $\mu_p = 2.792847350(9)\mu_N$. This measurement outperforms previous Penning-trap measurements[4,5] in terms of precision by a factor of about 760. It improves the precision of the forty-year-old indirect measurement, in which significant theoretical bound state corrections[6] were required to obtain μ_p, by a factor of 3. By application of this method to the antiproton magnetic moment, the fractional precision of the recently reported value[7] can be improved by a factor of at least 1,000. Combined with the present result, this will provide a stringent test of matter/antimatter symmetry with baryons[8].

The challenge to measure the properties of the proton with great precision inspires very different branches of physics. As part of the intense search for baryon number violation so far, a lower limit of the proton's lifetime of $t_p > 2.1 \times 10^{29}$ years has been set[9]. Employing Penning traps, the proton's atomic mass was measured with a fractional precision of 0.14 parts per billion (p.p.b.; ref. 10), and the proton-to-electron mass ratio was determined with a relative accuracy of 94 parts in 10^{12} (ref. 11). Both provide essential input parameters with which to test the theory of quantum electrodynamics and contribute to the search for physics beyond the Standard Model. Furthermore, exciting results obtained by spectroscopy of muonic hydrogen[12] yielded a new value of the proton charge radius, and compared to previous measurements a 7σ deviation was observed, which has yet to be understood.

Another property of the proton is its spin magnetic moment

$$\mu_p = g_p \frac{q_p}{2m_p} S \qquad (1)$$

where q_p/m_p is the charge-to-mass ratio and S is the particle's spin. The constant g_p is a dimensionless measure of μ_p in units of the nuclear magneton $\mu_N = q_p\hbar/(2m_p)$, where $\hbar$ is the reduced Planck constant. The most precise value of μ_p (see Fig. 1, white bar) is extracted from spectroscopy of atomic hydrogen[1]. In this experiment the bound proton-to-electron magnetic moment ratio $\mu_p(H)/\mu_e(H)$ was measured with a fractional precision of 10 p.p.b., and μ_p was calculated by taking theoretical corrections at the level of about 18 parts per million into account[6].

A scheme for the direct measurement of magnetic moments of single particles in Penning traps has been applied with great success in measurements of the $g-2$ values (where g is the dimensionless measure of the electron magnetic moment μ_e in units of the Bohr magneton $\mu_B = q_e\hbar/(2m_e)$) of the electron and the positron[13] and further improved for the electron in ref. 14, where fractional precisions at the level of 3.8 and 0.28 parts in 10^{12} were achieved, respectively. The application of this scheme to measure the magnetic moment of the proton is a considerable challenge, because μ_p is about 658 times smaller than the magnetic moment of the electron μ_e. Thus, an apparatus with much higher sensitivity to the magnetic moment is needed.

In a Penning trap, the g-factor of the proton is determined by the measurement of a frequency ratio $g_p/2 = \nu_L/\nu_c$ where $\nu_c = q_pB_0/(2\pi m_p)$ is the cyclotron frequency, and $\nu_L = (g_p/2)\nu_c$ is the spin-precession frequency, also called the Larmor frequency. Both frequencies are measured in the same magnetic field B_0. The cyclotron frequency is obtained by the Brown–Gabrielse invariance theorem, $\nu_c^2 = \nu_+^2 + \nu_z^2 + \nu_-^2$, where ν_+, ν_z and ν_- are the characteristic oscillation frequencies of the trapped particle[15], the modified cyclotron frequency, the axial frequency and the magnetron frequency, respectively. The Larmor frequency can be measured by application of the so-called continuous Stern–Gerlach-effect[3]. In that scheme a magnetic field inhomogeneity $\Delta B(z)_z = B_2z^2$ is superimposed on the trap, where B_2 characterizes its strength and z is the axial coordinate. This 'magnetic bottle' couples the spin magnetic moment to the axial oscillation frequency, thus reducing the determination of the spin state to a measurement of ν_z. A spin-flip shifts the axial frequency by $\Delta\nu_{z,SF} = (\mu_pB_2)/(2\pi^2m_p\nu_z)$. This enables the measurement of the spin transition rate as a function of an applied drive frequency ν_{rf}, and yields the Larmor frequency ν_L (ref. 16).

In our experiment, we use $B_2 = 2.97(10) \times 10^5\,\mathrm{T\,m^{-2}}$ (ref. 4), which is 2,000 times stronger than in the electron/positron experiments[13]. In the presence of such a strong magnetic inhomogeneity the axial frequency shift caused by a spin-transition is $\Delta\nu_{z,SF} = 171$ mHz out of $\nu_z \approx 740$ kHz. Thus, the detection of proton spin quantum jumps requires an adequate axial frequency stability that is difficult to achieve in the strong magnetic bottle B_2 (ref. 17). However, dramatic progress in the manipulation of a single trapped proton allowed the first direct measurements of the proton magnetic moment μ_p (refs 4 and 5; see Fig. 1, grey bars). Those experiments were carried out solely in Penning traps with a superimposed B_2, which are usually called 'analysis traps'. The strong inhomogeneity broadens the width of the spin resonance, and ultimately limits the precision to the level of parts per million. An elegant solution to boost experimental precision is provided by the double Penning-trap technique[2]. This method separates the analysis of the spin state from the precision measurements of ν_c and ν_L. In addition to the analysis trap, a precision trap is used, in which the magnetic field is more homogeneous by orders of magnitude. This narrows the width of the Larmor resonance dramatically, and thus improves the precision. Here we report the first direct measurement of the proton magnetic moment using this technique.

Figure 2a shows a schematic of our double Penning-trap setup located at the University of Mainz, Germany. It is mounted in the horizontal and southward-oriented bore of a superconducting magnet, with a magnetic field of $B_0 \approx 1.89$ T and a stability of $(\Delta B/B_0)/\Delta t = 4.0(0.7) \times 10^{-9}\,\mathrm{h^{-1}}$.

[1]Institut für Physik, Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany. [2]Helmholtz-Institut Mainz, 55099 Mainz, Germany. [3]RIKEN, Ulmer Initiative Research Unit, Wako, Saitama 351-0198, Japan. [4]Max-Planck-Institut für Kernphysik, 69117 Heidelberg, Germany. [5]Fakultät für Physik und Astronomie, Ruprecht-Karls-Universität Heidelberg, 69047 Heidelberg, Germany. [6]GSI—Helmholtzzentrum für Schwerionenforschung, 64291 Darmstadt, Germany. †Present address: RIKEN, Ulmer Initiative Research Unit, Wako, Saitama 351-0198, Japan.

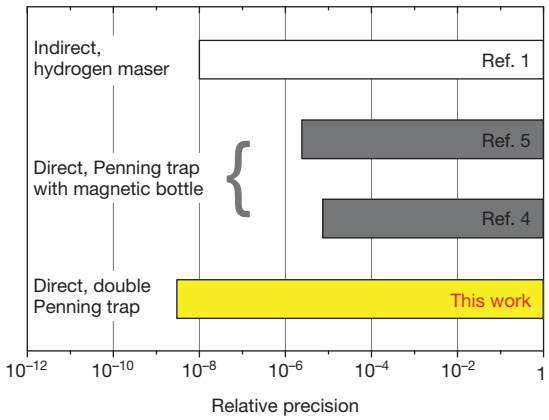


Figure 1 | Relative precision achieved in measurements of the proton magnetic moment. The value extracted indirectly from measurements with a hydrogen maser has a precision of 10 p.p.b. (ref. 1). Direct measurements with a single proton in a Penning trap with strong superimposed magnetic inhomogeneities were performed in 2012 by us[4] and a group at Harvard[5]. The result of the measurement described in this work was achieved by using the double Penning-trap technique with a single trapped proton. Our result is 3 times more precise than reported in ref. 1 and about 760 times more precise than other direct single particle measurements.

Each trap consists of five stacked cylindrical electrodes. The central ring electrode of the analysis trap is made out of ferromagnetic Co/Fe material, which generates the magnetic bottle. The other electrodes are made out of oxygen-free copper. To prevent oxidation all electrodes are gold-plated. The two trap centres are separated by 43.7 mm. Within this distance the magnetic-field inhomogeneity drops significantly. In the precision trap (PT) we have $B_{PT} = B_0 + B_{1,PT}z + B_{2,PT}z^2$, where $B_{1,PT} \approx 85\ \mathrm{T\,m^{-1}}$ and $B_{2,PT} \approx 4\ \mathrm{T\,m^{-2}}$, which is 75,000 times smaller than in the analysis trap. To shuttle the particle from one trap to the other, potential ramps are applied to transport electrodes located between the two traps. Radio-frequency drives applied to coils mounted close to each trap generate oscillating magnetic fields to drive proton spin transitions. The entire setup is placed in a hermetically sealed vacuum chamber cooled to 4 K. In this chamber pressures below 10^{-14} Pa are achieved[18], providing single-proton storage times of at least one year.

Protons are produced with an in-trap electron impact ion source. Electrons from a field emission gun hit a polyethylene target. Sputtered atoms are ionized in the centre of the precision trap. From the loaded ion-cloud a single proton is prepared using well established techniques[19].

Once a single proton is prepared, the particle's modified cyclotron mode is cooled resistively. This is achieved by tuning a cryogenic tank circuit, which acts on resonance as a resistor, to the cyclotron mode at $\nu_+ = 28.9$ MHz (ref. 20). Subsequently, the particle is shuttled to the analysis trap and the axial frequency is measured. To this end, $\nu_z \propto \sqrt{V_0}$ is tuned to resonance with our superconducting axial detection system at 740 kHz by adjusting the trapping voltage V_0. Once the axial motion is cooled to thermal equilibrium, the particle shorts the thermal noise of the detector and appears as a dip in the fast Fourier transform of the detector signal[21]. Such a fast Fourier transform spectrum is recorded in 60 s (shown as red points in Fig. 2b). By applying a fit to the data, the axial frequency ν_z is obtained. From this measurement, the quantum number of the cyclotron mode n_+ is determined[22]. Low n_+ is crucial to achieve an axial frequency stability which is sufficient to resolve single spin transitions[23,24]. Below a threshold cyclotron quantum number $n_{+,\mathrm{th}} < 1{,}500$ we achieve single spin-flip resolution with a fidelity

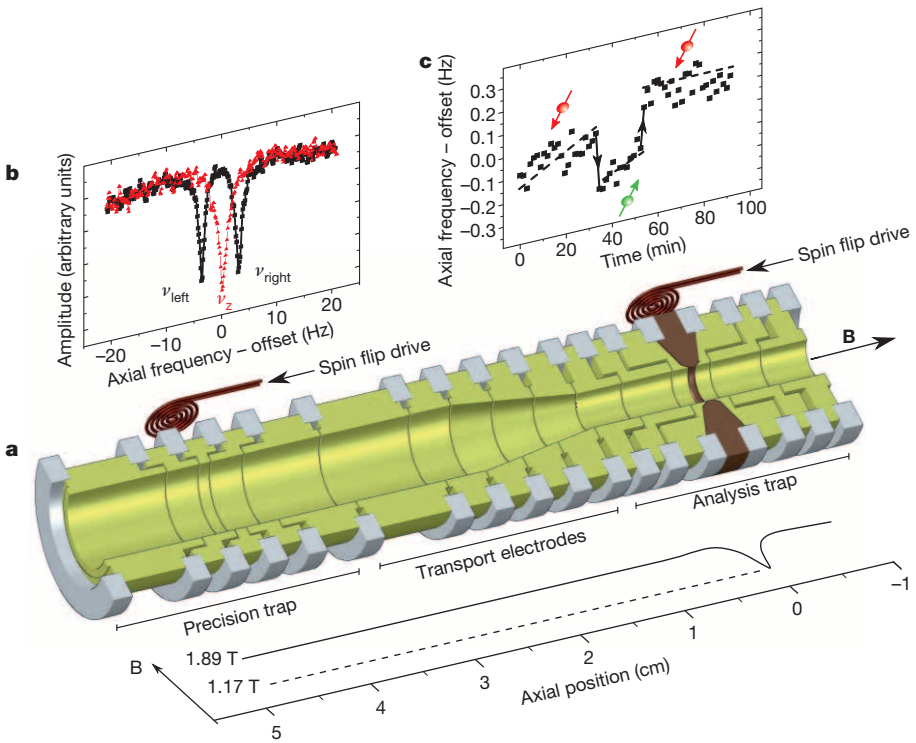


Figure 2 | Experimental setup and measurement procedures. a, Schematic of the double Penning-trap setup that is used for the direct measurement of the proton magnetic moment. It consists of two traps, an analysis trap and a precision trap, which are connected by transport electrodes. A strong magnetic field inhomogeneity is superimposed on the analysis trap, which is required to detect proton spin quantum transitions. In the precision trap, where the magnetic field is homogeneous, the precise frequency measurements are carried out. The solid curve in the plot below the trap system indicates the strength of the on-axis magnetic field. **b**, Fast Fourier transform spectrum of the axial detector's output signal. The dip (red) is due to a single particle, which shorts the thermal noise of the detector. The double-dip signal (black) appears when a quadrupolar sideband drive at $\nu_+ - \nu_z$ is applied. From such dip spectra ν_+, ν_z and ν_- are determined and thus the cyclotron frequency. The axial frequency has been offset by 623,850 Hz. For further details see text. **c**, Axial frequency measurement as a function of time. Frequency jumps of about 170 mHz are observed, which are due to induced single-proton spin transitions. The axial frequency has been offset by 742,060 Hz.

$F > 75\%$. This means that three out of four spin transitions are identified correctly. If $n_+ > n_{+,\mathrm{th}}$ is obtained, the particle is shuttled back to the precision trap and the cyclotron mode is cooled again. This procedure is repeated until $n_+ < n_{+,\mathrm{th}}$, which takes about two hours.

Once a particle with adequate n_+ is prepared, the actual g-factor measurement is conducted. First, the proton's spin state is identified. To this end, the axial frequency is measured and a spin-transition is induced by applying a magnetic radio-frequency drive, followed by another measurement of ν_z. As soon as a spin quantum jump is observed (see Fig. 2c), the proton's spin state is known. Afterwards, the particle is transported to the precision trap, where its cyclotron frequency ν_c is determined. First, the modified cyclotron frequency is measured via sideband coupling. An electric field at ν_{sb}, which is close to $\nu_+ - \nu_{z,\mathrm{PT}}$ is applied. This transfers energy between the axial mode and the modified cyclotron mode, and leads to a modulation of the axial oscillation amplitude. In the corresponding fast Fourier transform spectrum a so-called 'double-dip' with frequencies ν_{left} and ν_{right} for the left and the right dip, respectively, is observed (shown as black data points in Fig. 2b). Subsequently, the axial frequency is recorded, which is about $\nu_{z,\mathrm{PT}} \approx 624$ kHz. We determine ν_+ by applying the relation[25]

$$\nu_+ = \nu_{\mathrm{sb}} + \nu_{\mathrm{left}} + \nu_{\mathrm{right}} - \nu_{z,\mathrm{PT}} \quad (2)$$

The magnetron frequency $\nu_- \approx 7$ kHz is measured in a similar way. Finally, $\nu_{c,1}$ is obtained using the invariance theorem[15]. Next, spin transitions are induced by applying a spin-flip drive at $\nu_{\mathrm{rf,PT}}$ and subsequently the cyclotron frequency $\nu_{c,2}$ is measured again. Since the sideband drive leads to heating of the cyclotron mode, in a next step, by repeated cyclotron mode cooling in the precision trap and transport to the analysis trap, a low $n_+ < n_{+,\mathrm{th}}$ state is prepared and the spin state analysed again. From a comparison of the two measured spin states we conclude whether the spin in the precision trap was flipped. By repeating this scheme many times the spin-flip probability P_{SF} as a function of $\nu_{\mathrm{rf,PT}}$ is obtained. Normalizing each applied $\nu_{\mathrm{rf,PT},k}$ by the associated $\nu_{c,k}$ where k is the measurement number, one obtains a so-called g-factor resonance, $P_{SF}(\nu_L/\nu_c)$, with a maximum at $\mu_p/\mu_N = g_p/2$. For the normalization we use the average $(\nu_{c,1} + \nu_{c,2})/2$ of the two cyclotron frequency measurements. This compensates for linear magnetic field drifts which potentially take place while spin transitions are driven.

The result of our g-factor measurement is shown in Fig. 3. Incoherences caused by the coupling of the particle's axial motion to the axial detection circuit in the slightly inhomogeneous magnetic field of the precision trap prevent P_{SF} from exceeding 50% (ref. 16). The linewidth of the measured resonance is 12.5 p.p.b., which is caused by saturation and thermal fluctuations of the modified cyclotron energy E_+. The latter causes fluctuations of the axial frequency, $\nu_{z,\mathrm{PT}} \propto B_2 E_+$, which lead to fluctuations of the measured cyclotron frequency via equation (2), thus contributing to the linewidth. The measured data set is analysed using the maximum-likelihood method[26] based on a Gaussian distribution with the g-factor being the lineshape centre. The maximum-likelihood method is a statistical parameter estimation technique and avoids the need for arbitrary data binning, which may result in a biased estimate of the fitting parameters. From the data analysis we obtain

$$\mu_p = \frac{g_p}{2}\mu_N = 2.792\,847\,348(7)\mu_N$$

where the number in parentheses is the statistical uncertainty of the fit, which corresponds to a relative precision of 2.6 p.p.b.

Systematic shifts $\Delta(g_p/2)$ of the measured $(g_p/2)$ value are caused by time and energy dependencies of ν_L and ν_c

$$\frac{\Delta(g_p/2)}{(g_p/2)} = \frac{\Delta\nu_L(E_+,E_z,E_-,t)}{\nu_L} - \frac{\Delta\nu_c(E_+,E_z,E_-,t)}{\nu_c} \quad (3)$$

The frequency shifts $\Delta\nu_L(E_+, E_z, E_-, t)$ and $\Delta\nu_c(E_+, E_z, E_-, t)$ are caused by trap imperfections such as a slightly inhomogeneous magnetic field at the centre of the precision trap or an anharmonic trapping potential. To first order, the shifts induced by the magnetic field inhomogeneities cancel in the frequency ratio, because they contribute the same relative amount to ν_c and ν_L (refs 4 and 15). A considerable systematic shift can be caused by an anharmonicity of the electrostatic potential that only affects ν_c while ν_L remains unchanged. Thus, the trapping potential was optimized by careful adjustment of the voltages applied to the compensation electrodes of the precision trap to obtain $\Delta\nu_c = 0$. The relative uncertainty in the resulting shift $\Delta\nu_c$ is 0.20 p.p.b. A second g-factor resonance was recorded where the electrostatic potential was deliberately de-tuned to shift the modified cyclotron frequency by –5 p.p.b. Within error, we obtained the same $(g_p/2)$ value, which confirms that systematic shifts due to electrostatic anharmonicities are understood and negligible at the present level of precision. In addition to these two leading-order shifts, relativistic frequency shifts and image-charge shifts contribute[15]. However, because the frequency measurements are carried out in thermal equilibrium with the cryogenic detection system the relativistic shift contributes only at a level of 0.03 p.p.b. For our trap geometry the image-charge shift is −0.088 p.p.b. Thus, both shifts can be neglected. Time-dependent shifts are due to voltage and magnetic field drifts. The effect of voltage drifts was characterized by performing a sequence of axial frequency measurements at constant E_+. The corresponding systematic shift in ν_c is –0.07(0.35) p.p.b.

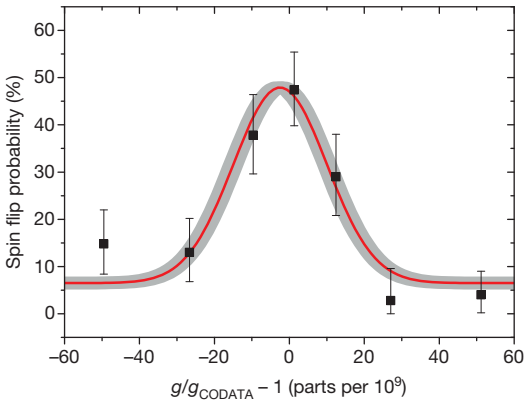


Figure 3 | Measured g-factor resonance. The abscissa is the measured g value normalized to the currently accepted value g_{CODATA}. The solid line is a maximum-likelihood fit to the data, which avoids the need for data binning. The shaded area indicates the 1σ confidence band of the fit. Filled squares representing binned data points with 1σ error bars are shown for visualization and do not explicitly enter the line fit. It took about four months to record the entire set of 450 data points.

The dominant systematic uncertainty is caused by nonlinear drifts of the magnetic field. By comparing the cyclotron frequency measurements before and after application of the spin-flip drive we find an average shift $\nu_{c,2} - \nu_{c,1} = 4$ p.p.b. Such frequency shifts in the precision trap are observed only if the spin-flip drive in the analysis trap has been applied in advance. It is consistent with heating of the electrodes caused by the latter spin-flip drive. Owing to thermal expansion the trap centres are shifted, thereby changing the magnetic field in the precision trap. Once the drive is turned off, thermal relaxation causes the observed drift of the magnetic field. The last contribution considered is a shift of the axial frequency after the measurement of ν_+. The sideband drive heats the particle to a cyclotron energy of $E_+/k_B = T_+ \approx 330$ K, where k_B is the Boltzmann constant. During the subsequent axial frequency measurement, the modified cyclotron mode is cooled resistively. This leads to an effective frequency shift of ν_z and contributes to a shift of the cyclotron frequency by –0.51(0.08) p.p.b. Accordingly, the magnetic moment value is corrected and the final result is

$$\frac{\mu_p}{\mu_N} = \frac{g_p}{2} = 2.792\,847\,350(7)(6) \quad (4)$$

The first and second numbers in parentheses are the statistical uncertainty of the fit and the systematic uncertainty, respectively, see Table 1. The latter is dominated by the nonlinear drift of the magnetic field. The result has a relative precision of 3.3 p.p.b. and is in agreement with the

Table 1 | Error budget of the direct proton *g*-factor measurement

Parameter	Relative shift of $g_p/2$	Uncertainty
Trapping potential	0	0.2×10^{-9}
Relativistic shift	0.03×10^{-9}	-
Image-charge shift	-0.088×10^{-9}	-
Voltage fluctuations	-0.07×10^{-9}	0.35×10^{-9}
Magnetic field relaxation	0	2×10^{-9}
Cyclotron cooling	-0.51×10^{-9}	0.08×10^{-9}
Total systematic shift	-0.64×10^{-9}	2×10^{-9}

This table lists the relative systematic shifts and their uncertainties, which were applied to the measured g_p value.

currently accepted CODATA value $g_{\mathrm{CODATA}}/2 = 2.792847356(23)$, but is 2.5 times more precise.

We expect it will be possible to achieve an improvement in precision by at least another factor of 10, by using an apparatus with reduced residual magnetic field inhomogeneity B_2 in the precision trap, a higher spin-state detection fidelity, as well as by applying phase-sensitive detection techniques[27]. In addition, the so-called Standard Model extension[8] describes diurnal frequency variations as a consequence of violation of the combined charge, parity and time symmetry and Lorentz violation. With faster measurement cycles, which become possible by application of advanced cyclotron cooling techniques, and an improved spin state detection fidelity a search for the predicted effects[28] becomes feasible.

The double Penning-trap method can be applied to measure the antiproton magnetic moment with similar precision[7,29]. A comparison of both values will provide a sensitive test of CPT invariance with baryons. The measurement of the antiproton magnetic moment will be conducted at the BASE experiment[30] at the Antiproton Decelerator of CERN.

Received 6 February; accepted 16 April 2014.

1. Winkler, P. F., Kleppner, D., Myint, T. & Walther, F. G. Magnetic moment of the proton in Bohr magnetons. *Phys. Rev. A* **5,** 83–114 (1972).
2. Häffner, H. *et al.* Double Penning-trap technique for precise *g*-factor determinations in highly charged ions. *Eur. Phys. J. D* **22,** 163–182 (2003).
3. Dehmelt, H. G. & Ekström, P. Proposed *g*-2 experiment on single stored electron or positron. *Bull. Am. Phys. Soc.* **18,** 727–731 (1973).
4. Rodegheri, C. C. *et al.* An experiment for the direct determination of the *g*-factor of a single proton in a Penning trap. *New J. Phys.* **14,** 063011 (2012).
5. DiSciacca, J. & Gabrielse, G. Direct measurement of the proton magnetic moment. *Phys. Rev. Lett.* **108,** 153001 (2012).
6. Karshenboim, S. G. & Ivanov, V. G. The *g*-factor of the proton. *Phys. Lett. B* **566,** 27–34 (2003).
7. DiSciacca, J. *et al.* One-particle measurement of the antiproton magnetic moment. *Phys. Rev. Lett.* **110,** 130801 (2013).
8. Bluhm, R., Kostelecký, V. A. & Russell, N. CPT and Lorentz tests in Penning traps. *Phys. Rev. D* **57,** 3932–3943 (1998).
9. Beringer, J. *et al.* (Particle Data Group). Review of particle physics. *Phys. Rev. D* **86,** 010001 (2012).
10. Van Dyck, R. S. Jr, Farnham, D. L., Zafonte, S. L. & Schwinberg, P. B. High precision Penning trap mass spectroscopy and a new measurement of the proton's "atomic mass". *AIP Conf. Proc.* **457,** 101–110 (1999).
11. Sturm, S. *et al.* High-precision measurement of the atomic mass of the electron. *Nature* **506,** 467–470 (2014).
12. Pohl, R. *et al.* The size of the proton. *Nature* **466,** 213–216 (2010).
13. Van Dyck, R. S., Schwinberg, P. B. & Dehmelt, H. G. New high-precision comparison of electron and positron *g*-factors. *Phys. Rev. Lett.* **59,** 26–29 (1987).
14. Hanneke, D., Fogwell, S. & Gabrielse, G. New measurement of the electron magnetic moment and the fine structure constant. *Phys. Rev. Lett.* **100,** 120801 (2008).
15. Brown, L. S. & Gabrielse, G. Geonium theory: physics of a single electron or ion in a Penning trap. *Rev. Mod. Phys.* **58,** 233–311 (1986).
16. Brown, L. S. Geonium lineshape. *Ann. Phys.* **159,** 62–98 (1985).
17. Ulmer, S. *et al.* Observation of spin flips with a single trapped proton. *Phys. Rev. Lett.* **106,** 253001 (2011).
18. Gabrielse, G. *et al.* Special relativity and the single antiproton: fortyfold improved comparison of $\bar{p}$ and p charge-to-mass ratios. *Phys. Rev. Lett.* **74,** 3544–3547 (1995).
19. Ulmer, S. *et al.* Direct measurement of the free cyclotron frequency of a single particle in a Penning trap. *Phys. Rev. Lett.* **107,** 103002 (2011).
20. Ulmer, S. *et al.* A cryogenic detection system at 28.9 MHz for the non-destructive observation of a single proton at low particle energy. *Nucl. Instrum. Meth. A* **705,** 55–60 (2013).
21. Wineland, D. J. & Dehmelt, H. G. Principles of the stored ion calorimeter. *J. Appl. Phys.* **46,** 919–930 (1975).
22. Mooser, A. *et al.* Demonstration of the double Penning-trap technique with a single proton. *Phys. Lett. B* **723,** 78–81 (2013).
23. Mooser, A. *et al.* Resolution of single spin-flips of a single proton. *Phys. Rev. Lett.* **110,** 140405 (2013).
24. DiSciacca, J., Marshall, M., Marable, K. & Gabrielse, G. Resolving an individual one-proton spin flip to determine a proton spin state. *Phys. Rev. Lett.* **110,** 140406 (2013).
25. Cornell, E. A., Weisskoff, R. M., Boyce, K. R. & Pritchard, D. E. Mode coupling in a Penning trap: π pulses and a classical avoided crossing. *Phys. Rev. A* **41,** 312–315 (1990).
26. Sivia, D. S. & Skilling J. *Data Analysis—A Bayesian Tutorial* 1st edn (Oxford Science, 2010).
27. Sturm, S. *et al. g*-factor measurement of hydrogenlike $^{28}Si^{13+}$ as a challenge to QED calculations. *Phys. Rev. A* **87,** 030501 (2013).
28. Mittleman, R. K., Ioannou, I. I., Dehmelt, H. G. & Russell, N. Bound on CPT and Lorentz symmetry with a trapped electron. *Phys. Rev. Lett.* **83,** 2116–2119 (1999).
29. Smorra, C. *et al.* Towards a high-precision measurement of the antiproton magnetic moment. *Hyperfine Interact.*. doi:10.1007/s10751-014-1018-7 (2014).
30. Ulmer, S. *et al. Technical Design Report BASE* CERN Document Server, SPSC-TDR-002, http://cds.cern.ch/record/1503514?ln=de (CERN, 2013).

Acknowledgements We acknowledge the financial support of the BMBF, and of the EU (ERC grant number 290870-MEFUCO), the Helmholtz-Gemeinschaft, HGS-HIRE, the Max-Planck Society, IMPRS-PTFS, and the RIKEN Initiative Research Unit Program.

Author Contributions S.U., C.C.R., H.K., A.M. and C.L. designed and built the experimental apparatus and the data acquisition system. A.M., C.L. and S.U. took part in the months-long data-taking runs. K.F. and A.M. developed the algorithms for the spin state analysis. A.M., K.F., S.U., C.L. and H.K. analysed the data. S.U., A.M., K.B. and J.W. wrote the initial manuscript, which was then improved and approved by all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.M. (mooser@uni-mainz.de).

LETTER

doi:10.1038/nature13376

Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle

Benjamin Poulter[1,2], David Frank[3,4], Philippe Ciais[2], Ranga B. Myneni[5], Niels Andela[6], Jian Bi[5], Gregoire Broquet[2], Josep G. Canadell[7], Frederic Chevallier[2], Yi Y. Liu[8], Steven W. Running[9], Stephen Sitch[10] & Guido R. van der Werf[6]

The land and ocean act as a sink for fossil-fuel emissions, thereby slowing the rise of atmospheric carbon dioxide concentrations[1]. Although the uptake of carbon by oceanic and terrestrial processes has kept pace with accelerating carbon dioxide emissions until now, atmospheric carbon dioxide concentrations exhibit a large variability on interannual timescales[2], considered to be driven primarily by terrestrial ecosystem processes dominated by tropical rainforests[3]. We use a terrestrial biogeochemical model, atmospheric carbon dioxide inversion and global carbon budget accounting methods to investigate the evolution of the terrestrial carbon sink over the past 30 years, with a focus on the underlying mechanisms responsible for the exceptionally large land carbon sink reported in 2011 (ref. 2). Here we show that our three terrestrial carbon sink estimates are in good agreement and support the finding of a 2011 record land carbon sink. Surprisingly, we find that the global carbon sink anomaly was driven by growth of semi-arid vegetation in the Southern Hemisphere, with almost 60 per cent of carbon uptake attributed to Australian ecosystems, where prevalent La Niña conditions caused up to six consecutive seasons of increased precipitation. In addition, since 1981, a six per cent expansion of vegetation cover over Australia was associated with a fourfold increase in the sensitivity of continental net carbon uptake to precipitation. Our findings suggest that the higher turnover rates of carbon pools in semi-arid biomes are an increasingly important driver of global carbon cycle inter-annual variability and that tropical rainforests may become less relevant drivers in the future. More research is needed to identify to what extent the carbon stocks accumulated during wet years are vulnerable to rapid decomposition or loss through fire in subsequent years.

Each year, on average, land and ocean carbon sinks absorb the equivalent of about half of the global fossil fuel emissions, thereby providing a critical service that slows the rise in atmospheric CO_2 concentrations[1]. Emissions from fossil fuels and land-use change now exceed ten billion tons or petagrams (Pg) of carbon per year, tracking the most carbon intense emission scenarios of the Intergovernmental Panel on Climate Change[4]. Even with this acceleration, the fraction of anthropogenic emissions that accumulates in the atmosphere (the airborne fraction) has remained largely unchanged since 1959 at 44% (ref. 2) ($P = 0.36$ for slope of linear regression). This implies that the uptake of carbon by ocean and terrestrial processes has, to some extent, kept pace with accelerating emissions owing to a range of possible factors, such as the fertilization effect of increased CO_2 and atmospheric nitrogen deposition on plant growth, changes in growing season length, and land management[5]. In addition to the continued uptake of CO_2, the airborne fraction exhibits large variability on interannual timescales, ranging between 18% and 79% during the past 54 years (ref. 2). This high interannual variability is primarily driven by terrestrial processes, which must be better understood in order for us to be able to forecast long-term biospheric responses to climate change[3].

High uncertainties in quantifying ecosystem processes mean that the global terrestrial carbon sink is often estimated as the residual between emissions from the combustion of fossil fuels, cement production and net land-use change, and sinks combining accumulation in the atmosphere and uptake by the ocean[6]. Using this method, the Global Carbon Project reported in their annual assessment a 2011 residual land sink of 4.1 Pg C yr^{-1} (standard deviation $\pm$ 0.9 Pg C yr^{-1}), representing an unusually large increase compared with the 2.6 $\pm$ 0.8 Pg C yr^{-1} decadal average and the largest reported residual land carbon sink since measurements of atmospheric CO_2 began in 1958. The 2011 residual land sink is indicative of several aspects of the debate surrounding the fate of terrestrial ecosystems under environmental change. First, the large uptake of carbon in 2011 continues a trend of increasing strength in the land carbon sink over at least one decade[1,7]. Second, the large annual growth anomaly in the land carbon sink raises questions regarding the growth rate of atmospheric CO_2 in coming years and how this is affected by the allocation of sequestered carbon to either labile or more stable pools. Lastly, increasing uncertainty in other terms of the global CO_2 budget has direct consequences on land sink estimates, for example, an overestimate of anthropogenic emissions would be assigned (owing to mass conservation and current accounting schemes) as an erroneously large land sink. Thus, attributing changes in net carbon uptake to carbon cycle processes requires a range of methodological approaches.

Here, we investigate the evolution of the terrestrial carbon sink over the past 30 years and the underlying mechanisms of the exceptionally large 2011 residual land carbon sink in a long-term context using (1) a 'bottom-up' process-oriented terrestrial biosphere model, (2) a 'top-down' atmospheric CO_2 inversion and (3) satellite observations of photosynthetic activity and vegetation structure. We allocate net land carbon uptake among specific geographic regions and provide a mechanistic explanation for the climatic and CO_2 response of net primary production (NPP), heterotrophic respiration (R_h), and disturbance that sum up to define net ecosystem exchange (NEE).

We find excellent agreement among the three different terrestrial carbon sink estimates that robustly support record 2011 land carbon uptake (Fig. 1a; with uncertainty presented as $\pm$1 standard deviation). The Lund–Potsdam–Jena (LPJ) dynamic global vegetation model (DGVM; ref. 8) estimates a 2011 land sink of 3.9 $\pm$ 1.3 Pg C yr^{-1}, a 1.3 $\pm$ 0.6 Pg C yr^{-1} anomaly compared to the 2003–2012 mean sink of 2.6 $\pm$ 0.9 Pg C yr^{-1} (Fig. 1a and Extended Data Table 1). Our atmospheric inversion (using the Monitoring Atmospheric Composition and Climate (MACC-II) inversion system; ref. 9) yields a 3.7 $\pm$ 0.4 Pg C yr^{-1} 2011 land sink, equivalent to a 1.0 Pg C yr^{-1} anomaly above the 2.7 $\pm$ 0.4 Pg C yr^{-1} inversion average for 2003–2012. The 2011 land sink estimates by the LPJ DGVM

[1]Montana State University, Institute on Ecosystems and the Department of Ecology, Bozeman, Montana 59717, USA. [2]Laboratoire des Sciences du Climat et de l'Environnement (LSCE), CEA CNRS UVSQ, 91191 Gif Sur Yvette, France. [3]Swiss Federal Research Institute WSL, Dendroclimatology, Zürcherstrasse 111, Birmensdorf 8903, Switzerland. [4]Oeschger Centre for Climate Change Research, University of Bern, CH-3012 Bern, Switzerland. [5]Department of Earth and Environment, Boston University, 685 Commonwealth Avenue, Boston, Massachusetts 02215, USA. [6]Faculty of Earth and Life Sciences, VU University Amsterdam, 1085 De Boelelaan, 1081HV, Amsterdam, The Netherlands. [7]Global Carbon Project, CSIRO, Marine and Atmospheric Research, Canberra, Australian Capital Territory 2601, Australia. [8]ARC Centre of Excellence for Climate Systems Science & Climate Change Research Centre, University of New South Wales, Sydney, New South Wales 2052, Australia. [9]Department of Ecosystem and Conservation Sciences, University of Montana, Missoula, Montana 59812, USA. [10]College of Engineering, Computing and Mathematics, University of Exeter, Exeter EX4 4QF, UK.

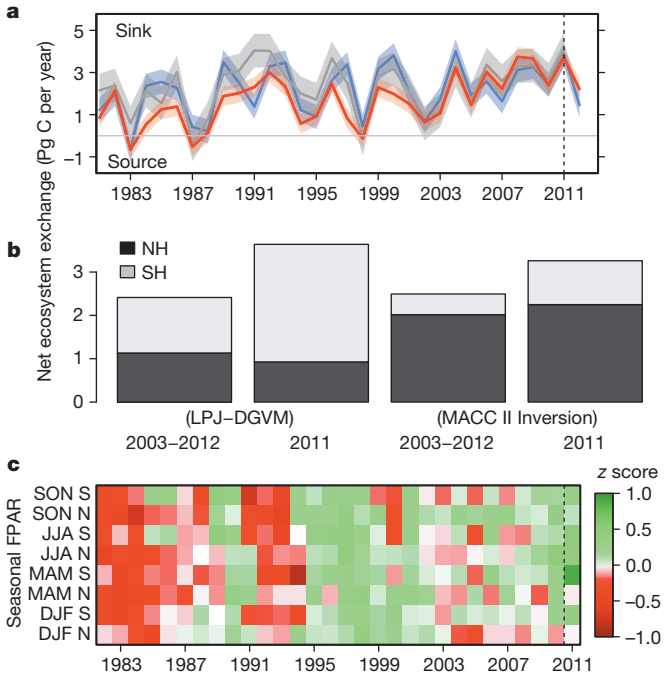


Figure 1 | Interannual variability of NEE and FPAR anomalies. **a**, Annual NEE, where positive values represent carbon uptake, blue is LPJ, red is MACC-II, and the residual land sink is in grey. The standard deviations are ± 0.58 Pg C yr^{-1} for LPJ, ± 0.4 Pg C yr^{-1} for the inversion, and ± 0.8 Pg C yr^{-1} for the residual (see Methods). **b**, Average, 2003–2012, annual NEE for the Northern and Southern hemispheres, estimated by LPJ and the inversion. **c**, AVHRR FPAR anomalies for the southern (S) and northern (N) hemispheres with respect to the 1982–2011 long-term average where the seasonal anomalies were calculated as the z score for each season (s) and each grid cell (i,j) for each year (y); $\mathrm{AVHRR}_{\mathrm{anomaly},ys(i,j)} = \dfrac{\mathrm{AVHRR}_{y,s(i,j)} - \mathrm{AVHRR}_{1982-2011,s(i,j)}}{\sigma \mathrm{AVHRR}_{1982-2011,s(i,j)}}$.

and MACC II inversion were greater than the 97.5th percentile over the period 1981–2012, suggesting a convergence of particularly novel ecosystem and climate states.

Both the atmospheric inversion and the DGVM demonstrate an increased contribution from Southern Hemisphere ecosystems to global net carbon uptake in 2011 (Fig. 1b). These patterns are supported by a large observed positive anomaly in the 2010–2011 interhemispheric CO_2 concentration gradient between Mauna Loa (19° N) and the Cape Grim (40° S) monitoring stations[10]. An increase in global NPP appears to be the main driving mechanism behind the 2011 land sink. Global NPP anomalies within the range of 1.7 Pg C simulated from the LPJ model forced with monthly climatic variables from the Climatic Research Unit (CRU) TS3.21 data set (ref. 11) and 1.6 Pg C by the Moderate Resolution Imaging Spectroradiometer (MODIS) NPP algorithm (Fig. 2a), using National Center for Environmental Prediction (NCEP) Reanalysis climate data and a light use-efficiency model[12] provide parallel support for this conclusion. Further investigation shows 79% (MODIS) to 87% (LPJ) of the global NPP anomaly is explained by just three semi-arid regions, Australia, temperate South America and southern Africa, where ecosystem respiration tends to lag productivity, inducing large net carbon uptake (regions AUST, SAmTe and SAf in Fig. 2b and Extended Data Fig. 1)[13–15]. In Australia, for example, compared with the 2003–2012 average, LPJ simulated a 45% increase in NPP for 2011, from an average of 1.75 to 2.54 Pg C yr^{-1}, but only a 9% increase in R_h (from 1.48 to 1.61 Pg C yr^{-1}). Moreover, wetter conditions decreased modelled fire-emissions by 29% (from 0.13 to 0.09 Pg C yr^{-1}) yielding a net 0.84 Pg C 2011 sink. Similarly, we find our conclusions for the greater sensitivity of NPP to precipitation, and lags in R_h, extend to southern Africa and temperate South America. In fact, 51% of the global 2011 net carbon sink was attributed to the three Southern Hemisphere semi-arid regions (Extended

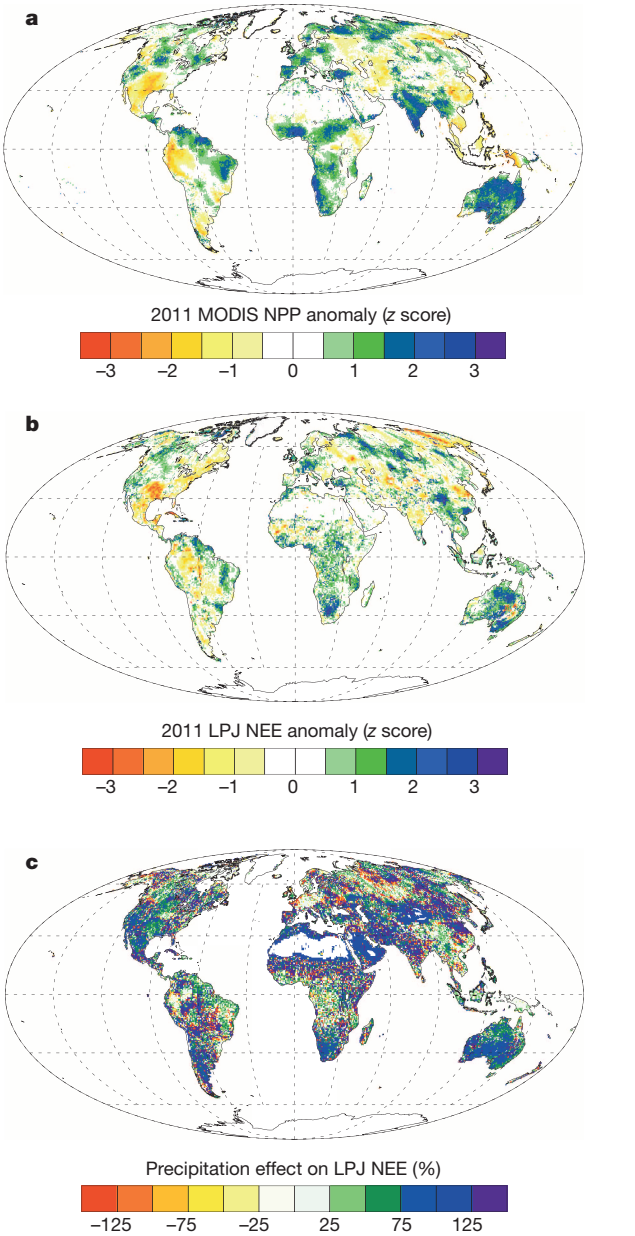


Figure 2 | Global anomalies of NPP and NEE, and the precipitation effect. **a**, Annual NPP anomaly, as z score (defined in Fig. 1), estimated by the MOD17A3 algorithm that uses the MODIS leaf area index (MOD15 Collection 5)[12]. **b**, Annual NEE anomaly, as z score, estimated by the LPJ DGVM, where a positive z score equals a larger sink; the reference period is 2000–2011. **c**, Spatial pattern of the contribution of precipitation to net ecosystem exchange in 2011 calculated as the difference between NEE with the all climate forcing varied and NEE simulated with the precipitation climatology (see Extended Data Fig. 6a and b for the NPP and R_h component fluxes).

Data Table 2), and Australia alone contributed to 57% of the total global LPJ NEE anomaly.

In addition to MODIS, the fraction of photosynthetic active radiation (FPAR) determined from the satellite-borne Advanced Very High Resolution Radiometer (AVHRR), AVHRR-FPAR3g (ref. 16), provides a long-term record of space-borne observations of the fraction of photosynthetic active radiation absorbed. Vegetation greening was widespread globally in 2011, with austral winter (June–August; JJA) FPAR reaching the highest values ever observed in the entire satellite period (1982–2011). In the Southern Hemisphere, record greening (Fig. 1c) was centralized over the same three Southern Hemisphere semi-arid regions (AUST, SAmTe and SAf) and was sustained for nine months spanning

2010 to 2011 (December–February, DJF; March–May, MAM; and JJA). Seasonal FPAR increases over Australia ranged from 4.6% in DJF and 8.7% in MAM to 5.1% in JJA, with all anomalies being prominent extremes in the context of an observed 0.8–1.9% interannual variability over the past 30 years. Notably, 46% (or 34%) of the land area in Australia experienced increases in FPAR in 2011 of more than 2.5 (or 3.0) standard deviations from normal in MAM, with positive FPAR anomalies first developing in eastern Australia in DJF, extending to all of Australia in MAM, then remaining in northern Australia in JJA (Extended Data Fig. 2).

To identify proximate causes for the role of semi-arid regions in the 2011 global sink, we performed a full set of LPJ factorial model simulations to isolate the temperature, precipitation, cloud cover and CO_2 contribution to NEE (Extended Data Table 1; Methods). An additional 'memory' simulation was conducted to evaluate previous-year climate effects that might have contributed to the extraordinary sink in 2011; the 2010 climate was replaced with a near-neutral year (2009) for the El Niño Southern Oscillation (Extended Data Fig. 3). With respect to pre-industrial CO_2 concentrations (287 parts per million, p.p.m.), the LPJ simulations suggest that CO_2 fertilization enhanced the 2011 net carbon uptake by 4.8 Pg C. High precipitation during 2010 and 2011 contributed to 0.62 Pg C and 0.52 Pg C of the global sink, respectively (Fig. 2c), or about 12%, thereby helping to offset land-to-atmosphere CO_2 fluxes driven by long-term negative temperature (−0.84 Pg C) and direct radiative contributions (−0.32 Pg C). In addition, 'memory' effects from 2010 added to the 2011 sink, with the largest difference being a threefold increase in tropical South American NEE when using 2009 climate before 2011. The increase in Amazonian NEE in 2011 was mainly due to recovery from the 2010 Amazon drought[17], which caused a reduction in LPJ NPP and an increase in LPJ R_h in 2010, leading to reduced short-lived litter carbon pools available for respiration and fire in 2011. Although 2011 precipitation explained most of the NEE increase in Australia (a 0.56 Pg C yr^{-1} contribution), the climate memory effect also explained 0.21 Pg C of the 2011 Australian sink because of high precipitation in 2010 that recharged soil moisture and plant carbohydrate reserves to the benefit of NPP in 2011. Among an ensemble of climate indices, the Multivariate El Niño Index (MEI; ref. 18) consistently explained the highest amount of year-to-year variability over Australia for annual carbon uptake (correlation coefficient $r = -0.49$, $P < 0.01$) and DJF FPAR greening ($r = -0.52$, $P < 0.01$) between 1981 and 2011 (Extended Data Fig. 4a–d). This extends earlier findings that found Pacific sea surface temperature as a significant predictor of precipitation-driven greening anomalies as far as South Africa and Australia[19,20]. Notably, the 2010/2011 La Niña—that is, the MEI negative phase—took place over an especially long time period, as observed from multiple satellite, rain gauge and reanalysis data sources (the Tropical Rainfall Measurement Mission (TRMM), CRU and NCEP-DOE; Extended Data Fig. 5a and b), and even lowered global sea levels[21], in addition to altering global carbon uptake[12].

The available evidence points towards an enhanced climatic effect of the 2010/2011 La Niña from interactions with long-term semi-arid region greening trends since the early 1980s at the latest. For example, since 1982, we found an expansion of vegetation across the Australian landscape ($P < 0.01$ for one-sided Kolmogorov–Smirnov test) where land area with FPAR >20% (or 30%) increased by 5.6% (or 3.5%) in the MAM growing season. The greening trend in semi-arid regions has been previously associated with a range of drivers that include altered precipitation frequency and intensity[22], increased water-use efficiency due to elevated CO_2 effects on leaf stomatal conductance[23], and woody encroachment following land-use and grazing[22,24]. Over the same 1982–2011 time period, we observed a statistically significant increase in the sensitivity of LPJ net carbon uptake ($P < 0.001$) and AVHRR-FPAR3g vegetation activity ($P < 0.02$) to austral-autumn precipitation for the Australian continent (Fig. 3a). The observed change in ecosystem sensitivity over Australia meant that an additional 100 mm of growing season (MAM) precipitation led to a fourfold increase in net carbon uptake when comparing sensitivities before (0.2 Pg C yr^{-1} per 100 mm) or after (0.8 Pg C yr^{-1} per 100 mm) 1997, the midpoint of current observational records (1982–2011). An independent data-driven model of net ecosystem production[25], which excluded disturbance processes, confirmed the same statistically robust increase over time in carbon uptake per unit precipitation for Australia (Fig. 3b, $P < 0.001$). Long-term observations from passive-microwave vegetation optical depth (VOD)[26] suggest that the enhanced sensitivity of vegetation to climate is a result of both increases in grass cover as well as from woody encroachment (Fig. 3c).

The 2011 land carbon sink anomaly indicates a novel climate-driven response of the biosphere where interactions between extremes in austral precipitation[27] and changes in land cover[23] (both possibly caused by humans) are contributing to non-analogue ecosystem behaviour with global biogeochemical significance. We propose that the current paradigm, whereby tropical rainforest El Nino/Southern Oscillation coupling dominates interannual variability of the atmospheric CO_2 growth rate[3,28], may become less relevant in the future. We explored whether semi-arid carbon-cycle climate sensitivity feedbacks exist among an ensemble of 15 Earth system models contributed to the Coupled Model Intercomparison Project Phase 5 (CMIP5; ref. 29). In contrast to our observations, we found that for semi-arid regions, modelled carbon uptake and precipitation sensitivity remains relatively stable over the 1990 to 2090 period for the CMIP5 ensemble ($P = 0.33$, one-sided Student's t-test, Fig. 4). This suggests that processes contributing to the novel ecosystem dynamics identified here may be overlooked in future climate change scenarios. As the dynamics of dryland systems, which cover 45% of the Earth's land surface, increase in global importance, more

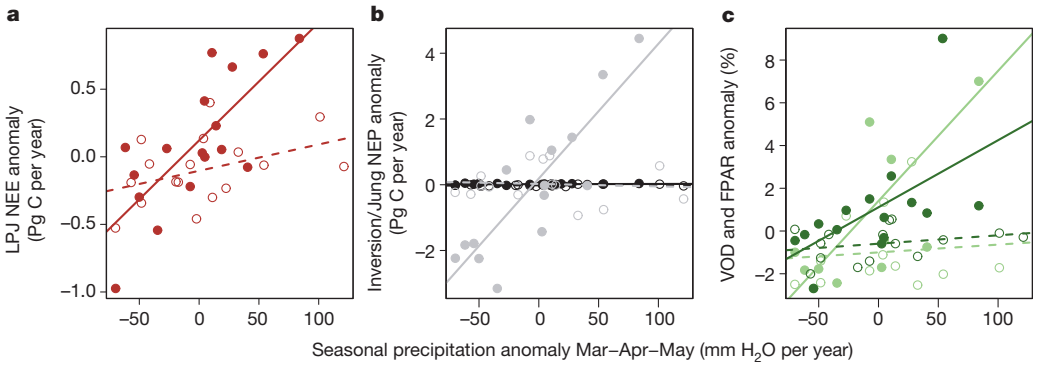


Figure 3 | Change in climate sensitivity of observations for Australia. **a**, Climate sensitivity of annual LPJ NEE anomalies to MAM precipitation anomalies for Australia. The open circles and dashed line are the points and regression line for 1982–1996 (β_1) and the closed circles and solid line are for 1997–2011 ($\beta_1 + \beta_3$), from the following linear regression model using NEE and precipitation anomalies ($P_{anomaly}$) where A is an indicator variable for the different time periods: ($NEE_{anomaly} = \beta_0 + \beta_1 P_{anomaly} + \beta_2 A + \beta_3 P_{anomaly} A$). **b**, Climate sensitivity of annual NEE from the MACC-II inversion (black symbols) and the upscaled NEP product (grey symbols) using the same linear model as in **a**. **c**, Climate sensitivity of annual VOD (light green symbols) and MAM FPAR (dark green symbols) also using the model described in **a**.

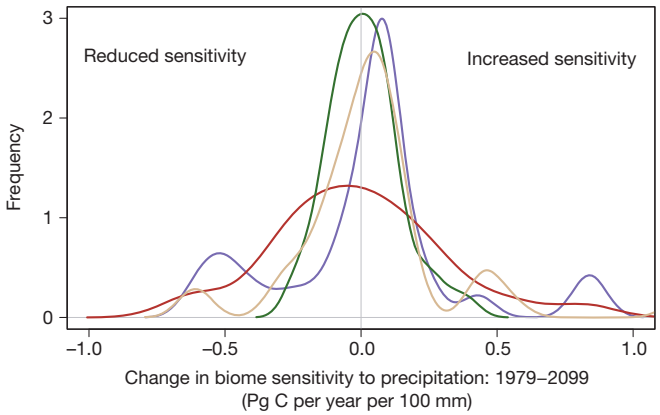


Figure 4 | Change in regional climate sensitivity of CMIP5 models. Distribution of the change in sensitivity between the periods 1979–2005 and 2069–2095 in net biome production to annual precipitation for four biomes ($n = 15$ CMIP5 Earth System Models). Precipitation sensitivity was estimated as β_1 while controlling for changes in sensitivity due to CO_2 and air temperature $NBP_{anomaly} = \beta_0 + \beta_1 CO_2^{anomaly} + \beta_2 P^{anomaly} + \beta_3 T^{air}_{anomaly}$. The different lines refer to tropical (green), temperate (brown), semi-arid (tan), and boreal (purple) biomes.

research is needed to identify whether enhanced carbon sequestration in wet years is particularly vulnerable to rapid decomposition or loss through fire in subsequent years, and is thus largely transitory. Such behaviour may already be reflected by the larger-than-average atmospheric growth rate in 2012 (ref. 30) that was associated with a return to near-normal terrestrial land sink conditions (Fig. 1a).

METHODS SUMMARY

We use multiple data sources, including carbon accounting methods, carbon-cycle model simulations, and satellite-based vegetation products to investigate the magnitude and mechanisms driving variability in the terrestrial carbon sink. NPP (the total photosynthesis minus plant autotrophic respiration losses) is simulated by the LPJ DGVM and also estimated independently with the MODIS NPP algorithm, MOD17A3. The balance between carbon uptake from NPP and losses from soil respiration and disturbance (NEE) is quantified from the Global Carbon Project, the LPJ DGVM, and the MACC-II atmospheric inversion system (http://www.copernicus-atmosphere.eu/). NEP (the balance between gross carbon inputs from photosynthesis and losses from ecosystem respiration, excluding disturbance) is estimated from upscaled gridded flux tower observations. Optical and passive microwave satellite data are employed to assess vegetation greenness trends (AVHRR FPAR3g) and vegetation structure or vegetation optical depth (VOD). Monthly and seasonal precipitation fluctuation is quantified from TRMM 3B43v7 (http://mirador.gsfc.nasa.gov) and NCEP-DOE Reanalysis II (http://www.esrl.noaa.gov), and the Climatic Research Unit (CRU) TS3.21 (http://www.cru.uea.ac.uk/). Regional summaries of the global gridded data followed boundaries from the eleven land regions specified in the TRANSCOM atmospheric inversion experiment. We further differentiate North and South Africa to distinguish between wet and semi-arid climates with the ratio of precipitation to potential evaporation set to 0.7. Historical (1860–2005) simulations of net biome production, equivalent to NEE, from the Fifth Coupled Model Intercomparison Project (CMIP5) are merged with the Representative Concentration Pathway 8.5 (RCP8.5) to create temporal composites spanning 1860–2099 for 15 Earth system models.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 August 2013; accepted 10 April 2014.
Published online 21 May 2014.

1. Ballantyne, A. P., Alden, C. B., Miller, J. B., Tans, P. P. & White, J. W. C. Increase in observed net carbon dioxide uptake by land and oceans during the past 50 years. *Nature* **488,** 70–72 (2012).
2. Le Quéré, C. *et al.* The global carbon budget 1959–2011. *Earth Syst. Sci. Data* **5,** 1107–1157 (2013).
3. Cox, P. *et al.* Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature* **494,** 341–344 (2013).
4. Peters, G. P. *et al.* The challenge to keep global warming below 2 °C. *Nature Clim. Change* **3,** 4–6 (2012).
5. Pan, Y. *et al.* A large and persistent carbon sink in the world's forests. *Science* **333,** 988–993 (2011).
6. Canadell, J. G. *et al.* Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks. *Proc. Natl Acad. Sci. USA* **104,** 18866–18870 (2007).
7. Sitch, S. *et al.* Trends and drivers of regional sources and sinks of carbon dioxide over the past two decades. *Biogeosci. Disc.* **10,** 20113–20177 (2013).
8. Sitch, S. *et al.* Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Glob. Change Biol.* **9,** 161–185 (2003).
9. Chevallier, F. *et al.* CO_2 surface fluxes at grid point scale estimated from a global 21 year reanalysis of atmospheric measurements. *J. Geophys. Res. D* **115,** D21307 (2010).
10. Francey, R. J. *et al.* Atmospheric verification of anthropogenic CO_2 emission trends. *Nature Clim. Change* **3,** 520–524 (2013).
11. Harris, I., Jones, P. D., Osborn, T. J. & Lister, D. H. Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. *Int. J. Climatol.* **34,** 623–642 (2013).
12. Bastos, A., Running, S. W., Gouveia, C. & Trigo, R. M. The global NPP dependence on ENSO: La-Niña and the extraordinary year of 2011. *J. Geophys. Res.* **118,** 1247–1255 (2013).
13. Haverd, V. *et al.* Multiple observation types reduce uncertainty in Australia's terrestrial carbon and water cycles. *Biogeosciences* **10,** 2011–2040 (2013).
14. Haverd, V. *et al.* The Australian terrestrial carbon budget. *Biogeosciences* **10,** 851–869 (2013).
15. Rotenberg, E. & Yakir, D. Contribution of semi-arid forests to the climate system. *Science* **327,** 451–454 (2010).
16. Zhu, Z. *et al.* Global Data Sets of Vegetation LAI3g and FPAR3g derived from GIMMS NDVI3g for the period 1981 to 2011. *Remote Sens.* **5,** 927–948 (2013).
17. Marengo, J. A., Tomasella, J., Alves, L. M., Soares, W. R. & Rodriguez, D. A. The drought of 2010 in the context of historical droughts in the Amazon region. *Geophys. Res. Lett.* **38,** L12703 (2011).
18. Wolter, K. & Timlin, M. S. Monitoring ENSO in COADS with a seasonally adjusted principal component index. In *Proc. 17th Climate Diagnostics Workshop* 52–57 (NOAA/NMC/CAC, NSSL, Univ. Oklahoma, 1993).
19. Myneni, R. B., Los, S. O. & Tucker, C. J. Satellite-based identification of linked vegetation index and sea surface temperate anomaly areas from 1982–1990 for Africa, Australia and South America. *Geophys. Res. Lett.* **23,** 729–732 (1996).
20. Woodward, F. I., Lomas, M. R. & Quaife, T. Global responses of terrestrial productivity to contemporary climatic oscillations. *Phil. Trans. R. Soc. Lond. B* **363,** 2779–2785 (2008).
21. Boening, C., Willis, J. K. Landerer, F. W., Nerem, R. S. & Fasullo, J. The 2011 La Niña: so strong, the oceans fell. *Geophys. Res. Lett.* **39,** doi:10.1029/2012GL053055 (2012).
22. Donohue, R. J., McVicar, T. R. & Roderick, M. L. Climate-related trends in Australian vegetation cover as inferred from satellite observations, 1981–2006. *Glob. Change Biol.* **15,** 1025–1039 (2009).
23. Donohue, R. J., Roderick, M. L., McVicar, T. R. & Farquhar, G. D. CO_2 fertilisation has increased maximum foliage cover across the globe's warm, arid environments. *Geophys. Res. Lett.* **40,** 3031–3035 (2013).
24. Asner, G. P., Elmore, A. J., Olander, L. P., Martin, R. E. & Harris, A. T. Grazing systems, ecosystem responses, and global change. *Annu. Rev. Environ. Resour.* **29,** 261–299 (2004).
25. Jung, M., Reichstein, M. & Bondeau, A. Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. *Biogeosciences* **6,** 5271–5304 (2009).
26. Andela, N., Liu, Y. Y., van Dijk, A. I. J. M., de Jeu, R. A. M. & McVicar, T. R. Global changes in dryland vegetation dynamics (1988–2008) assessed by satellite remote sensing: combining a new passive microwave vegetation density record with reflective greenness data. *Biogeosci.* **10,** 6657–6676 (2013).
27. Kang, S. M. *et al.* Modeling evidence that ozone depletion has impacted extreme precipitation in the austral summer. *Geophys. Res. Lett.* **40,** 4054–4059 (2013).
28. Wang, W. *et al.* Variations in atmospheric CO_2 growth rates controlled by tropical temperature. *Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.1219683110 (2013).
29. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of the CMIP5 and the experimental design. *Bull. Am. Meteorol. Soc.* **93,** 485–498 (2012).
30. Le Quéré, C. *et al.* Global carbon budget 2013. *Earth Syst. Sci. Data Discuss.* **6,** 689–760 (2013).

Acknowledgements We acknowledge support from the EU FP7 GEOCARBON programme (283080), and thank the researchers involved with collecting and maintaining the climate data at the Climate Research Unit, University of East Anglia, UK, and the National Center for Atmospheric Research, USA. We also acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP. We thank the climate modelling groups for producing and making available their model output. For CMIP the US Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We also thank M. Jung for providing the 'upscaled' NEE data used in our analysis. J.G.C. acknowledges the support of the Australian Climate Change Science Program. R.B.M. and S.W.R. were funded by the NASA Earth Science Division. C. Le Quéré and L. Cernusak provided comments and suggestions that improved the manuscript. This paper is a contribution to the Global Carbon Budget activity of the Global Carbon Project.

Author Contributions B.P., D.F., P.C. and R.B.M. designed the analyses. J.B., F.C., G.B., D.F., R.B.M., S.W.R., S.S., G.R.v.d.W., J.G.C., Y.Y.L. and N.A. contributed data to the analyses. B.P., F.C., R.B.M., S.R. and D.F. conducted the analyses. All authors contributed to the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.P. (benjamin.poulter@montana.edu).

LETTER

doi:10.1038/nature13262

Storm-induced sea-ice breakup and the implications for ice extent

A. L. Kohout[1], M. J. M. Williams[2], S. M. Dean[2] & M. H. Meylan[3]

The propagation of large, storm-generated waves through sea ice has so far not been measured, limiting our understanding of how ocean waves break sea ice. Without improved knowledge of ice breakup, we are unable to understand recent changes, or predict future changes, in Arctic and Antarctic sea ice. Here we show that storm-generated ocean waves propagating through Antarctic sea ice are able to transport enough energy to break sea ice hundreds of kilometres from the ice edge. Our results, which are based on concurrent observations at multiple locations, establish that large waves break sea ice much farther from the ice edge than would be predicted by the commonly assumed exponential decay[1–3]. We observed the wave height decay to be almost linear for large waves—those with a significant wave height greater than three metres—and to be exponential only for small waves. This implies a more prominent role for large ocean waves in sea-ice breakup and retreat than previously thought. We examine the wider relevance of this by comparing observed Antarctic sea-ice edge positions with changes in modelled significant wave heights for the Southern Ocean between 1997 and 2009, and find that the retreat and expansion of the sea-ice edge correlate with mean significant wave height increases and decreases, respectively. This includes capturing the spatial variability in sea-ice trends found in the Ross and Amundsen–Bellingshausen seas. Climate models fail to capture recent changes in sea ice in both polar regions[4,5]. Our results suggest that the incorporation of explicit or parameterized interactions between ocean waves and sea ice may resolve this problem.

Sea ice is a feature of both polar regions and has an important role in moderating the global climate. The expansion and contraction of sea ice is largely governed by seasonal changes in air temperature, with the finer details controlled by the complex feedbacks that exist between sea ice, the atmosphere and the ocean. Of these feedbacks, the interaction between ocean waves and sea ice is one of the least well understood and is usually overlooked in coupled climate models. Yet the ability of waves to break sea ice has been known since the 'heroic age' of polar exploration[6]. Waves propagating through sea ice leave behind a wake of broken ice floes, which are then more easily deformed by winds and currents, effectively eliminating the barrier between air and ocean and enhancing heat exchange.

The key to predicting the magnitude of ice breakup lies in understanding wave attenuation in the marginal ice zone (MIZ), which is a region, potentially hundreds of kilometres wide, of broken ice floes that forms at the boundary of the open ocean and the sea ice at each pole. Present models of the breakup process either show only moderate agreement with measurements[7], or results are not compared to measured data[8]. Models have typically depended on measurements collected during the 1970s and early 1980s[9,10], in experiments conducted over short timescales and in relatively low-amplitude ocean swells. Large storm waves are routinely found in the Southern Ocean[11] and may be anticipated in an increasingly ice-free Arctic. These drive large-wave events that have occasionally been observed at single locations within the MIZ[12–14], but these observations are insufficient to determine how storm-generated waves propagate through sea ice.

Here we report the measurement of wave attenuation using simultaneous observations across hundreds of kilometres in the Antarctic MIZ, and examine its implications for the recent retreat and expansion of Antarctic sea ice. Five wave sensors were deployed on sea ice between latitudes 60.5° south and 63° south on 23 and 24 September 2012 UTC (Fig. 1). Along the deployment transect, the average ice floe diameter increased steadily from 2–3 m at the ice edge to 10–20 m approximately 200 km from the ice edge. Beyond this, there was an abrupt increase in floe diameter to hundreds of metres (Extended Data Table 1). Ice was estimated from manual shipboard observations to be between 0.5 and 1 m thick and was all first-year ice. The rate at which sea-ice concentration increased with distance from the edge was high relative to the climatological rate for this location (Extended Data Fig. 1). The significant wave heights measured by the sensors include relatively calm conditions and three large-wave events (Extended Data Fig. 2). On 1 October 2012 UTC significant wave heights of 3 m were measured 240 km from the ice edge.

Analysis of wave decay in sea-ice focuses on understanding the evolution of the full wave spectrum propagating through the ice. Linear theory assumes that as a wave propagates through ice, the power at each wavenumber decays without transfer of energy between wave numbers. This implies that the significant wave height, which is proportional to the

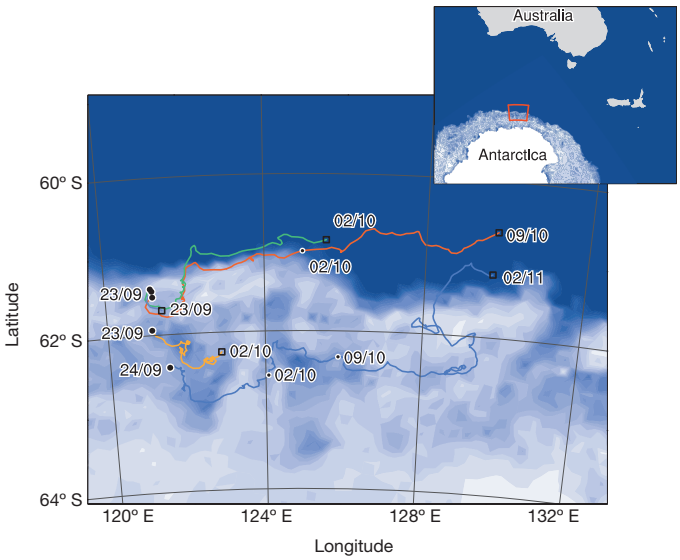


Figure 1 | Deployment location and track of each wave sensor. The large round markers show where and when (day and month UTC) each sensor was deployed. The open square markers show where and when each sensor stopped transmitting. The small round markers indicate the sensor positions on particular dates. Inset, location of the experiment on a larger scale, the red box indicating the main figure. Mean sea-ice concentrations between 23 September and 2 October 2012 are shown with white as 100% sea-ice concentration and blue as open water.

[1]National Institute of Water and Atmospheric Research, Christchurch 8011, New Zealand. [2]National Institute of Water and Atmospheric Research, Wellington 6021, New Zealand. [3]The University of Newcastle, Newcastle, New South Wales 2308, Australia.

square root of the total wave energy, will always decay exponentially with distance from the sea-ice edge[14,15]. Our results confirm previous observations[9,10,16] that, during calm conditions, the significant wave height decays exponentially with distance. However, during three large-wave events, we found that significant wave heights did not decay exponentially, enabling large waves to persist deep into the pack ice.

To demonstrate this, we calculate the decay rate of the significant wave height between wave buoys, dH_s/dx, where H_s is the significant wave height and x the distance between buoys. Using observations farther than 100 km from the ice edge, the magnitude of dH_s/dx increases almost perfectly linearly with H_s until H_s reaches 3 m (Fig. 2). For waves larger than 3 m, dH_s/dx flattens and can be treated as being independent of H_s. This shows that existing linear theory is only valid for waves with $H_s < 3$ m. The empirical model derived from the data is

$$\frac{dH_s}{dx} = \begin{cases} -5.35 \times 10^{-6} H_s & H_s \leq 3 \\ -16.05 \times 10^{-6} & H_s \geq 3 \end{cases} \quad (1)$$

where -5.35×10^{-6} is the attenuation coefficient. The attenuation coefficients estimated from observations of small waves in the Arctic[10] are also shown in Fig. 2, for comparison. The constant attenuation for waves with significant wave height greater than 3 m implies a more gradual decay of wave height with propagation distance, allowing large waves to penetrate considerably farther into the ice. Because the ice in the MIZ was all first-year ice, we are unable to determine how equation (1) will differ in thicker ice or in a combination of first-year and multi-year ice.

The wave spectra during large-wave events (Extended Data Fig. 3) indicate that the spectral peak of the energy distribution may shift to longer periods with increasing distance from the ice edge. This is standard for waves in the open ocean, where nonlinear interactions create an inverse energy cascade, moving energy and the spectral peak to longer periods[17]. Thus, our observations suggest that nonlinear energy transfer may need to be considered when modelling the decay of large waves ($H_s > 3$ m) through sea ice and that small-amplitude wave theory cannot simply be extrapolated to large-amplitude waves.

We use our new model for wave decay (equation (1)) to estimate the distance from the ice edge over which a wave will be able to break ice floes. Following classical strain theory and using minimum and maximum breaking strains of 3×10^{-5} and 7.05×10^{-5}, respectively[7,8], we consider a wave with a 12-s period travelling through first-year sea ice, and predict ice breaking as a function of significant wave height at the ice edge, assuming that $H \approx H_s/\sqrt{2}$, where H is the single-period wave height (Fig. 3 and Methods). This model can be extended to consider a wave spectrum[8], but the results are almost identical to our single-period analysis. This figure expresses how the wave attenuation that we have observed during large-wave events has a profound impact by transferring energy deep into the ice pack. Also shown are two known observations of ice floe breakup events. One occurred during this experiment (in the Antarctic) and the other occurred in the Arctic[18]. Each observation is a minimum bound on the distance the ice broke during the respective storms. Figure 3 also shows the breaking limit calculated using the Arctic observations[10] and the consequence of extrapolating small-amplitude wave theory to large-amplitude waves.

Motivated by our experimental results, we examine the wider relevance of wave breakup on sea-ice extent, proposing that an increase in significant wave height in the Southern Ocean would increase the breakup of sea ice, resulting in a retreat of the sea-ice edge. Conversely, for a decrease in significant wave height we would expect the ice edge to expand. To identify whether such a relationship exists, we compare model estimates of significant wave heights[19] with satellite sea-ice observations between 1997 and 2009, taking the 15% ice concentration contour to define the sea-ice edge. The data set was divided into two six-month seasons per year: growth (March to August) and decay (September to February) (Fig. 4). Both show that decreasing and increasing trends in sea-ice extent correlate with increasing and, respectively, decreasing trends in significant wave height (Pearson correlation coefficients were −0.70 in the decay season and −0.79 in the growth season). We find that a 2-m increase in significant wave height over a decade leads to a 2° latitudinal retreat in

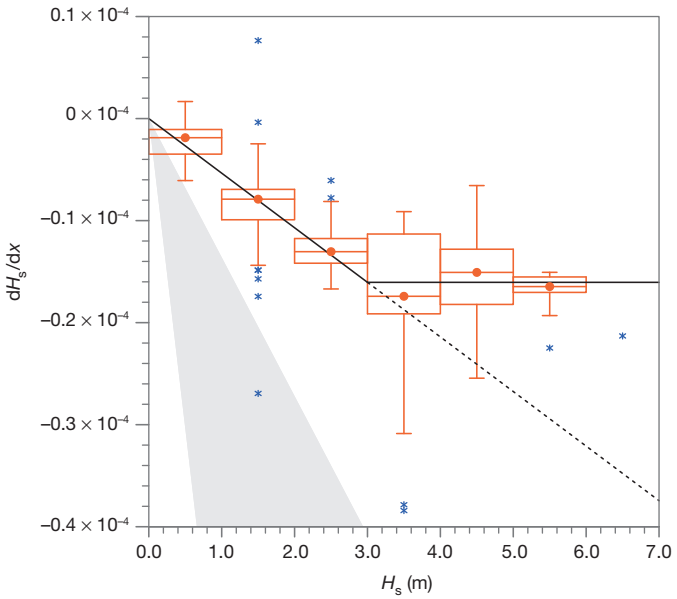


Figure 2 | Decay rates of sensors farther than 100 km from the ice edge. Data are binned in 1-m boxes. The red dot is the median. Box height shows the range within which 50% of the data lie. The whiskers give the range of data, excluding outliers (blue markers) and single data points. The solid black line is calculated from linear least-squares regression through the median values. The dashed black line shows the decay that would be expected if small-amplitude wave theory held for large waves. The grey region gives the range of decay rates observed in the Arctic[10].

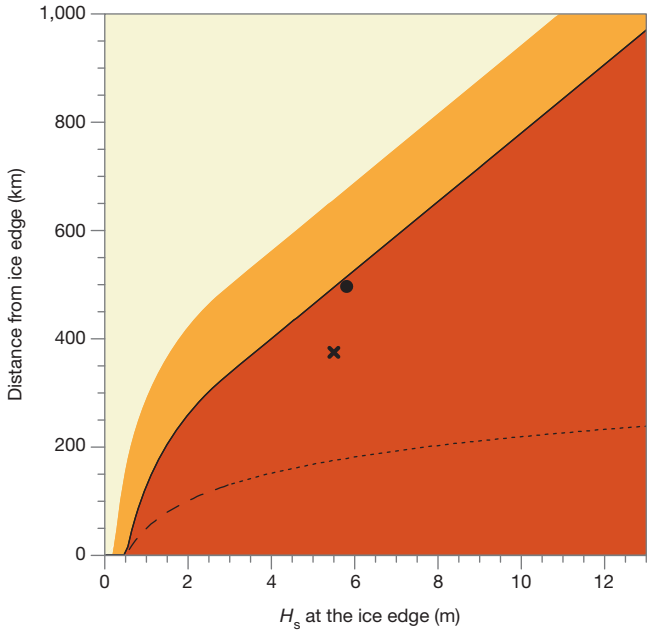


Figure 3 | Ice-breaking potential as a function of the distance from the ice edge and the significant wave height at the ice edge. Highlighted are conditions where breaking will not occur (yellow), may occur (orange) and is likely to occur (red). The long-dashed line is the likely breaking limit calculated using the attenuation coefficient observed in the Arctic[10] for small-amplitude waves with a 12-s period. This is extrapolated using small-amplitude wave theory (short-dashed). The markers show ice floe breakup events during this experiment at 11:00 on 1 October 2012 UTC (dot) and during a large-wave event in the Arctic[18] (cross).

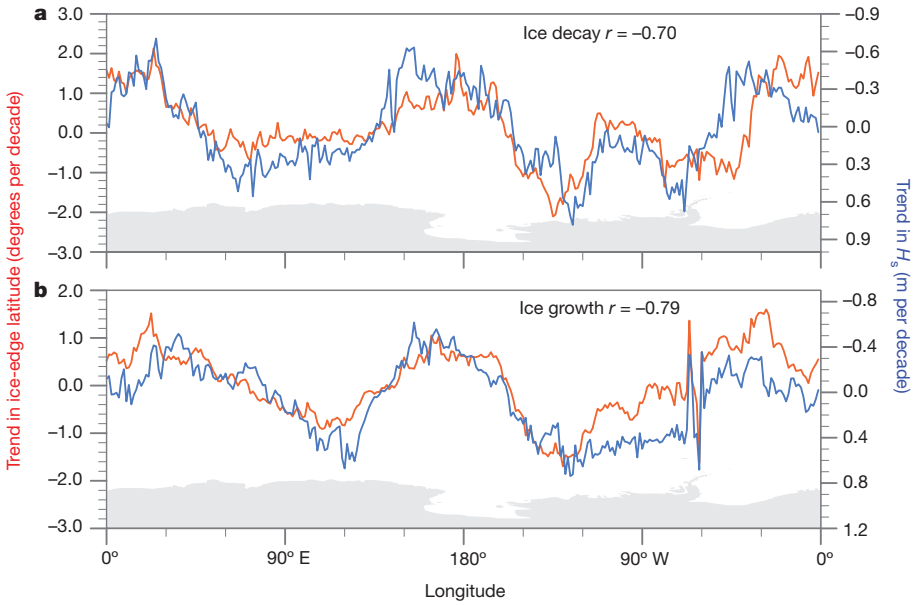


Figure 4 | A comparison between the trends in sea-ice extent and significant wave height between 1997 and 2009. The observed trend in the location of the ice edge (red) and the simulated trend in the significant wave height (blue) are shown as functions of longitude. **a**, Averaged trends for the ice decay season (September to February). **b**, The averaged trends during the ice growth season (March to August). The Pearson coefficient (r) is given at the top right of each panel ($n = 288$ for each). Antarctica is represented by the grey shaded region. We note that the scale for trend in H_s increases downwards.

ice extent (Extended Data Fig. 4). Spatially, the largest increases in wave height were found in the Amundsen–Bellingshausen Sea and the largest decreases were found in the Western Ross Sea (Fig. 4), areas where regional sea-ice retreat and expansion, respectively, are well documented[20]. The trends in significant wave height between 1997 and 2009 are consistent with the observed longer-term trends[11]. Identifying a similar relationship in the Arctic was not possible, owing to insufficient data.

Our results suggest that sea ice is vulnerable to changes in storminess. The observed southward shift in the storm tracks over recent decades has resulted in fewer cyclones at mid latitudes and more cyclones at higher latitudes[21]. However, the response has not been purely zonal; recent trends in the surface pressure pattern and winds have exhibited zonally asymmetric changes, resulting in variability in the atmospheric forcing of sea ice[22]. In the future, wave heights are predicted to increase everywhere at the sea-ice edge in the Arctic and Antarctic[23]. It is conceivable that this will act to accelerate sea-ice retreat.

Climate models continue to have difficulty in accurately predicting both Arctic sea-ice retreat and the regional variations in sea ice around Antarctica, suggesting either the inaccurate representation or the omission of important sea-ice physics. Our new observations show that slower decay by large waves (>3 m), possibly owing to nonlinear processes, allows waves to maintain ice-breaking potential hundreds of kilometres into the pack ice. This is a relationship that seems to be consistent around Antarctica, with changing wave heights correlating with changes in the latitude of the ice edge. This suggests that wave/sea-ice interactions need to be included in climate models, either directly or as parameterizations, before these models can improve their representations of sea ice. Although some sea-ice models are beginning to consider wave/sea-ice interactions, they may have limited success if they inadequately represent the behaviour of large storm-driven waves.

METHODS SUMMARY

The waves-in-ice observation system consisted of a high-resolution accelerometer coupled with a tri-axis inertial measurement unit (IMU), which was located using the Global Positioning System. Instruments simultaneously recorded wave accelerations for 34 min every 3 h. This record was filtered and integrated to calculate displacement, and a subsampled fast Fourier transform of the data was returned via satellite. Almost 600 data records were returned[24] (Extended Data Table 2). The data set is complex, owing to the significant spatial and temporal variation in wave forcing. To quantify how waves decay, we assume that the wave field is consistent along the zonal spread of the sensors, and that as waves enter the sea ice, they refract and travel south along a meridian. We compare only simultaneously collected instrument records, under the assumption that the wave climate persisted long enough to ensure that the wave spectra captured deep within the ice pack are from the wavefront captured near the ice edge. We further integrate the data by considering the decay of total wave energy, that is, the significant wave height, rather than the decay at individual frequencies. For each adjacent pair of sensors, we calculate the decay in significant wave height as a function of distance. The pairs are grouped into 1-m bins, determined by the significant wave height of the northernmost sensor. To take advantage of sea ice acting as a low-pass filter, we consider only pairs of sensors whose mean distance from the ice edge is greater than 100 km (Extended Data Table 3). Full details on our experiment design and analysis techniques are provided in Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 December 2013; accepted 14 March 2014.

1. Wadhams, P. *Ice in the Ocean* 202–218 (Gordon and Breach, 2002).
2. Squire, V. A., Dugan, J. P., Wadhams, P., Rottier, P. J. & Liu, A. K. Of ocean waves and sea ice. *Annu. Rev. Fluid Mech.* **27,** 115–168 (1995).
3. Dumont, D., Kohout, A. & Bertino, L. A wave-based model for the marginal ice zone including a floe breaking parameterization. *J. Geophys. Res.* **116,** C04001 (2011).
4. Stroeve, J., Holland, M. M., Meier, W., Scambos, T. & Serreze, M. Arctic sea ice decline: faster than forecast. *Geophys. Res. Lett.* **34,** L09501 (2007).
5. Turner, J., Bracegirdle, T. J., Phillips, T., Marshall, G. J. & Hosking, J. S. An initial assessment of Antarctic sea ice extent in the CMIP5 models. *J. Clim.* **26,** 1473–1484 (2013).
6. Shackleton, E. H. *South: The Story of Shackleton's Last Expedition, 1914–17* 103 (Macmillan, 1982).
7. Kohout, A. L. & Meylan, M. H. An elastic plate model for wave attenuation and ice floe breaking in the marginal ice zone. *J. Geophys. Res.* **113,** C09016 (2008).
8. Williams, T. D., Bennetts, L. G., Squire, V. A., Dumont, D. & Bertino, L. Wave-ice interactions in the marginal ice zone. Part 2: numerical implementation and sensitivity studies along 1D transects of the ocean surface. *Ocean Model.* **71,** 92–101 (2013).
9. Wadhams, P., Squire, V. A., Goodman, D. J., Cowan, A. M. & Moore, S. C. The attenuation rates of ocean waves in the marginal ice zone. *J. Geophys. Res.* **93,** 6799–6818 (1988).
10. Squire, V. A. & Moore, S. C. Direct measurement of the attenuation of ocean waves by pack ice. *Nature* **283,** 365–368 (1980).
11. Young, I. R., Zeiger, S. & Babanin, A. V. Global trends in wind speed and wave height. *Science* **332,** 451–455 (2011).
12. Doble, M. J. & Bidlot, J.-R. Wave buoy measurements at the Antarctic sea ice edge compared with an enhanced ECMWF WAM: Progress towards global waves-in-ice modelling. *Ocean Model.* **70,** 166–173 (2013).

13. Marko, J. R. Observations and analyses of an intense waves-in-ice event in the Sea of Okhotsk. *J. Geophys. Res.* **108,** C9 (2003).
14. Squire, V. A. Of ocean waves and sea-ice revisited. *Cold Reg. Sci. Technol.* **49,** 110–133 (2007).
15. Longuet-Higgins, M. S. On the statistical distributions of sea waves. *J. Mar. Res.* **11,** 245–265 (1952).
16. Liu, A. K., Vachon, P. W., Peng, C. Y. & Bhogal, A. S. Wave attenuation in the marginal ice zone during LIMEX. *Atmos.-ocean* **30,** 192–206 (1992).
17. Komen, G. J. *et al. Dynamics and Modelling of Ocean Waves* 232 (Cambridge Univ. Press, 1994).
18. Prinsenberg, S. J. & Peterson, I. K. Observing regional-scale pack-ice decay processes with helicopter-borne sensors and moored upward-looking sonars. *Ann. Glaciol.* **52,** 35–42 (2011).
19. Chawla, A., Spindler, D. M. & Tolman, H. L. Validation of a thirty year wave hindcast using the Climate Forecast System Reanalysis winds. *Ocean Model.* **70,** 189–206 (2013).
20. Parkinson, C. L. & Cavalieri, D. J. Antarctic sea ice variability and trends, 1979–2010. *Cryosphere* **6,** 871–880 (2012).
21. Hartmann, D. L. *et al.* in *Climate Change 2013: The Physical Science Basis* (eds Stocker, F. F. *et al.*) 159–254 (Cambridge Univ. Press, 2013).
22. Holland, P. R. & Kwok, R. Wind-driven tends in Antarctic sea-ice drift. *Nature Geosci.* **5,** 872–875 (2012).
23. Dobrynin, M., Murawski, J. & Yang, S. Evolution of the global wind wave climate in CMIP5 experiments. *Geophys. Res. Lett.* **39,** L18606 (2012).
24. Kohout, A. L. & Williams, M. J. M. Waves-in-ice observations made during the SIPEX II voyage of the Aurora Australis, 2012. *Australian Antarctic Data Centre* http://dx.doi.org/10.4225/15/53266BEC9607F (2013).

Acknowledgements We thank Inprod Pty Ltd for instrument design and construction; M. Doble, V. Squire and T. Haskell for contributions toward instrument design; T. Toyota for the provision of ice floe size and ice thickness data; the captain and crew of RSV *Aurora Australis* and the second Sea Ice Physics and Ecosystem Experiment (SIPEX II) for their assistance in deploying the waves-in-ice instruments; and V. Squire, T. Toyota, L. Bennetts and T. Williams for contributions toward interpretation and editing. The work was funded by a New Zealand Foundation of Research Science and Technology Postdoctoral award to A.L.K.; the Marsden Fund Council, administered by the Royal Society of New Zealand; NIWA, through core funding under the National Climate Centre Climate Systems programme; the Antarctic Climate and Ecosystems Cooperative Research Centre; and Australian Antarctic Science project 4073.

Author Contributions A.L.K. and M.J.M.W. had the idea for and designed the study. A.L.K. carried out the experiment. A.L.K., S.M.D. and M.H.M. analysed the data. All authors contributed to data interpretation and writing.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.L.K. (alison.kohout@niwa.co.nz).

LETTER

doi:10.1038/nature13195

A Palaeozoic shark with osteichthyan-like branchial arches

Alan Pradel[1], John G. Maisey[1], Paul Tafforeau[2], Royal H. Mapes[3] & Jon Mallatt[4]

The evolution of serially arranged, jointed endoskeletal supports internal to the gills—the visceral branchial arches—represents one of the key events in early jawed vertebrate (gnathostome) history, because it provided the morphological basis for the subsequent evolution of jaws[1–5]. However, until now little was known about visceral arches in early gnathostomes[6–17], and theories about gill arch evolution were driven by information gleaned mostly from both modern cartilaginous (chondrichthyan) and bony (osteichthyan) fishes. New fossil discoveries can profoundly affect our understanding of evolutionary history, by revealing hitherto unseen combinations of primitive and derived characters[18,19]. Here we describe a 325 million year (Myr)-old Palaeozoic shark-like fossil that represents, to our knowledge, the earliest identified chondrichthyan in which the complete gill skeleton is three-dimensionally preserved in its natural position. Its visceral arch arrangement is remarkably osteichthyan-like, suggesting that this may represent the common ancestral condition for crown gnathostomes. Our findings thus reinterpret the polarity of some arch features of the crown jawed vertebrates and invert the classic hypothesis, in which modern sharks retain the ancestral condition[3,20]. This study underscores the importance of early chondrichthyans in resolving the evolutionary history of jawed vertebrates.

The visceral skeleton of jawed vertebrates consists of a series of jointed arches including the jaws, hyoid arch and gill arches. This fundamental arrangement is shared by all chondrichthyans (sharks, rays and chimaeroids) and osteichthyans (bony fishes and their limbed relatives), as well as the extinct 'placoderm' and 'acanthodian' fishes. It thus represents a highly conserved feature of gnathostomes. Important differences in arch structure between modern chondrichthyans and osteichthyans[5,9,21,22] could reflect their long independent evolutionary history (over 420 Myr). The modern osteichthyan arrangement is already recognizable in early actinopterygian ('ray-finned') and sarcopterygian ('lobe-finned') osteichthyans (for example, *Pteronisculus*[8], *Mimipiscis*[9], *Eusthenopteron*[12]) and is commonly assumed to be derived, whereas the modern shark pattern is viewed as primitive[3,20]. Crucially, little was known about visceral arches in Palaeozoic chondrichthyans[11,13–15,17] (Supplementary Notes), precluding a detailed comparison with other early jawed vertebrates.

Articulated, three-dimensionally preserved specimens of a small symmoriiform shark (stem chondrichthyan[23]) from the Lower Carboniferous period of Arkansas, United States, were investigated by propagation phase contrast X-ray synchrotron microtomography, which revealed the complete series of visceral arches on both sides (Fig. 1 and Supplementary Video 1). Although several specimens were examined, the most complete example (American Museum of Natural History (AMNH) FF20544) is used here to illustrate our findings (Fig. 1 and Extended Data Fig. 1). The specimens possess tessellated calcified cartilage, which is considered to be the hallmark character of 'conventionally defined' chondrichthyans[19], a group that we informally name 'euchondrichthyans'.

Class Chondrichthyes Huxley, 1880
Order Symmoriiformes Zangerl, 1981
Family Falcatidae Zangerl, 1990
Ozarcus mapesae gen. et sp. nov.

Etymology. The generic name derives from Ozark (a highland region of Arkansas where the specimens were found) and ultimately from *arcus* (Latin for arch). The species is named after G. K. Mapes in recognition of her work collecting and describing fossils from Palaeozoic strata in the United States for more than 40 years, and who found the holotype specimen.

Holotype. AMNH FF20544 (Fig. 1 and Extended Data Fig. 1).

Referred material. Articulated heads: AMNH FF20525, 20528 and 20542.

Locality and horizon. Fayetteville Formation (lower shale member), Chesterian, Upper Mississippian, from the ARC-07 locality[24] (section (sec.) 22, township (T.) 14 N, range (R.) 15 W) near Leslie, Searcy County, Arkansas, United States.

Diagnosis. Falcatidae possessing small cladodont, pentacuspid and symmetrical teeth; ten upper and lower tooth families; palatoquadrate lacks a continuous posterior quadrate margin; no dermal denticles covering the top of the head; no 'spine–brush' complex; no laterally extended supraorbital shelf; laterally extended antorbital process that overlies a suborbital process possessing a series of four ridges and grooves for the articulation with the palatoquadrate (Supplementary Notes and Extended Data Fig. 2).

The visceral skeleton of *Ozarcus* has the same basic organization as in most other piscine jawed vertebrates, with a series of paired, jointed arches including a mandibular arch (jaw), hyoid arch (behind the jaw) and five branchial arches (Fig. 1). The palatoquadrate articulates with the postorbital process of the braincase (Fig. 1b) as in many other Palaeozoic sharks. No traces of labial cartilages were identified.

The hyoid arch of *Ozarcus* includes paired epihyals and ceratohyals, but neither pharyngohyals nor interhyals. Unlike in modern chondrichthyans[22], anteriorly directed, osteichthyan-like[21] paired hypohyals are present instead of a median basihyal (Figs 1f, 2 and Extended Data Fig. 3). Some other Palaeozoic euchondrichthyans also possess hypohyals, but the distribution of basihyals and hypohyals among most early gnathostomes is poorly resolved (Supplementary Notes).

The epi- and ceratohyals are morphologically similar to the corresponding branchial elements (Figs 1b, e, 2a), as in modern chimaeroids but unlike in elasmobranchs, where the hyoid arch is specialized towards jaw suspension[5]. The epihyal extends forward to just behind the orbit, but both the epihyal and the braincase lack any articular processes or recesses, suggesting that the epihyal was weakly connected to the braincase (Extended Data Fig. 3b, c). The epihyal meets the ceratohyal well posterior to the jaw joint, leaving a space between the mandibular and hyoid arches that, even allowing for compaction of the fossil, is far wider than those between the branchial arches (Fig. 1b and Extended Data Fig. 3a); a similar space has been inferred in the symmoriiform *Cobelodus*[15]. Its conjunction with the recessed palatoquadrate margin in *Ozarcus* may indicate the presence of a larger version of the non-respiratory, pseudobranch-bearing

[1]Department of Vertebrate Paleontology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024, USA. [2]European Synchrotron Radiation Facility, BP 220, 6 rue Jules Horowitz, 38043 Grenoble Cedex, France. [3]Department of Geological Sciences, Ohio University, Athens, Ohio 45701, USA. [4]School of Biological Sciences, Washington State University, Pullman, Washington 99164-4236, USA.

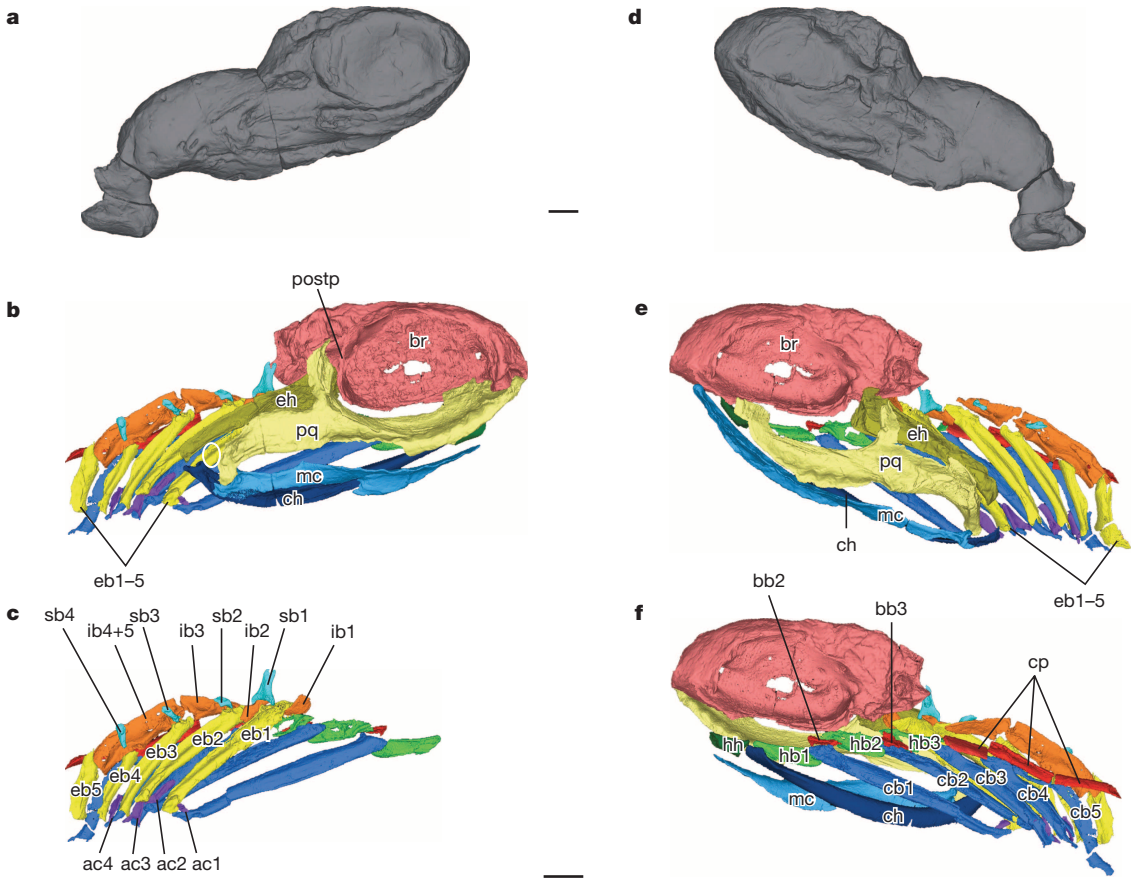


Figure 1 | Three-dimensional reconstructions of *Ozarcus mapesae* AMNH FF20544. **a**, Concretion in right lateral view. **b**, Braincase and associated visceral skeleton in right lateral view. **c**, Branchial skeleton in right lateral view (mandibular and hyoid arches removed). **d**, Concretion in left lateral view. **e**, Braincase and associated visceral skeleton in left lateral view. **f**, Braincase and the right elements of the visceral skeleton in medial aspect, left lateral view. Colour coding of the skeletal elements: yellow, epi-; blue, cerato-; green, hypo-; orange, infrapharyngo-; turquoise, suprapharyngo-; purple, accessory elements; red, basi-; peachy pink, braincase. The colours of the mandibular elements are lightened and those of the hyoid are darkened. The white circle indicates the space between the mandibular and hyoid arches. ac1–4, accessory elements of branchial arches 1–4; bb2–3, basibranchials 2–3; br, braincase; cb1–5, ceratobranchials 1–5; cp, copula of posterior basibranchials; eb1–5, epibranchials 1–5; eh, epihyal; hb1–3, hypobranchials 1–3; hh, hypohyal; ib1–4, infrapharyngobranchials 1–5; mc, Meckel's cartilage/lower jaw; postp, postorbital process; pq, palatoquadrate/upper jaw; sb1–4, suprapharyngobranchials 1–4. Scale bar, 10 mm.

spiracular pouch seen in modern jawed vertebrates[5], or perhaps even a fully respiratory hyomandibular gill pouch[1,3,16,25] (Supplementary Notes and Extended Data Fig. 4).

The branchial arches include the basi-, hypo-, cerato-, epi- and pharyngobranchials (Figs 1, 2), arranged in a way not previously observed in a euchondrichthyan. Instead of a single pharyngobranchial as in modern chondrichthyans, arches 1–4 each have two pharyngobranchials (Figs 1c, 2). We regard these as homologues of the infra- and suprapharyngobranchials in osteichthyans[6,21] (Supplementary Notes). The single pharyngobranchial in modern chondrichthyans is topographically homologous with osteichthyan infrapharyngobranchials[6,10], not with the suprapharyngobranchials[26], because both chondrichthyan pharyngobranchials and osteichthyan infrapharyngobranchials[6] support the roof of the pharynx. Unlike in modern chondrichthyans, the infrapharyngobranchials of *Ozarcus* are directed anteriorly, as in osteichthyans.

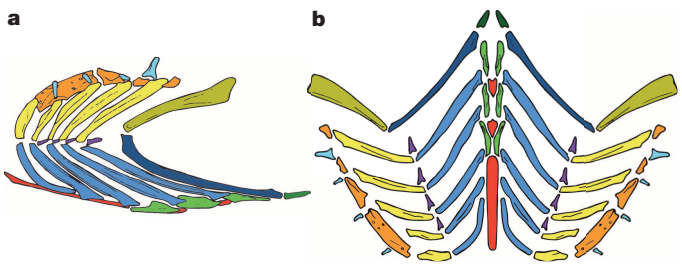


Figure 2 | Reconstructions of the branchial skeleton of *O. mapesae*. **a**, Right elements of the branchial skeleton (mandibular arch removed) in lateral aspect, right lateral view. **b**, Branchial skeleton reconstructed as horizontally spread, viewed dorsally and with anterior above (mandibular arch removed). Same colours and abbreviations as in Fig. 1. Not to scale.

A separate, fifth epibranchial element is present in *Ozarcus* (Figs 1b, c, e, 2), as well as in *Debeerius*[13], some extinct hybodont sharks (for example, *Tribodus*[27]) and *Acanthodes*[28]. By contrast, epibranchial element 5 is fused with pharyngobranchial 4 in modern chondrichthyans and is absent in crown osteichthyans[20].

The last ceratobranchial in *Ozarcus* is not broadened laterally, a feature shared with xenacanths, *Debeerius*[13] and chimaeroids. By contrast, in many modern sharks and the Carboniferous shark *Tristychius*[17], the last ceratobranchial is much wider than those farther anteriorly.

Ventrally, arches 3–5 meet the lateral margins of the basibranchial copula, as in osteichthyans (Figs 1f, 2), whereas only the last ceratobranchial is attached to the copula in modern chondrichthyans. All the hypobranchials are directed anteriorly in *Ozarcus* (Figs 1c, f, 2), an arrangement shared with osteichthyans[5], *Acanthodes*[10], modern chimaeroids[7] and *Debeerius*[13]. By contrast, only hypobranchial 1 is anteriorly directed in some modern elasmobranchs (for example, *Scyliorhinus canicula*; Fig. 3).

Small 'accessory cartilages' are present between the ceratobranchial and epibranchial on each of the first four gill arches (Figs 1c, e, 2). These were previously unknown in chondrichthyans, but similar elements are

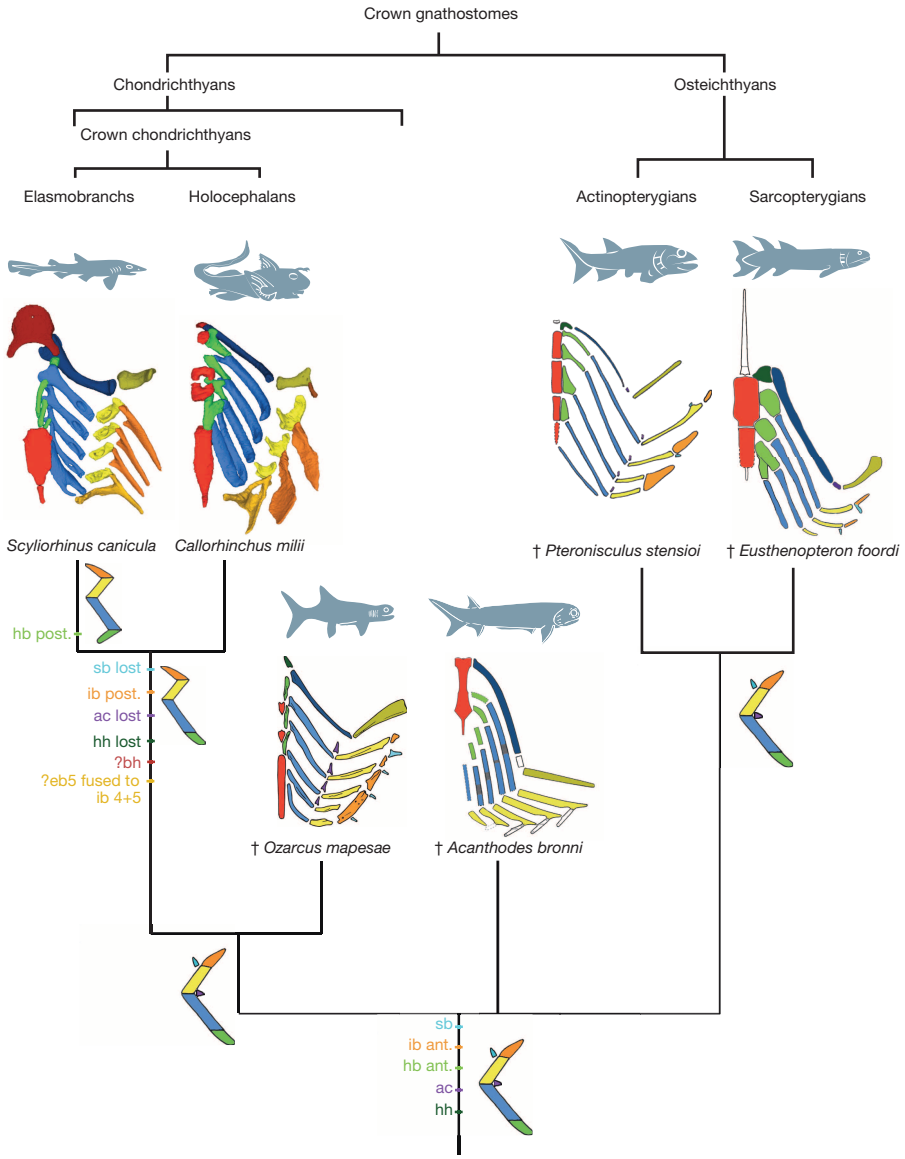


Figure 3 | Evolution of the branchial skeleton in the crown gnathostomes, mapped onto a tree compiled from the most recent phylogenetic analyses[19,23,28]. Only the right and median parts of the branchial skeletons of *Scyliorhinus*, *Callorhinchus*, *Ozarcus*, *Acanthodes*[10,28], *Pteronisculus*[8] and *Eusthenopteron*[12] are shown, as viewed dorsally and with anterior above. Diagrammatic right views of one branchial arch are shown at the crown gnathostome, chondrichthyan and osteichthyan nodes of the tree. To maximize clarity, the branchial skeletons are reconstructed as horizontally spread and the opercular cartilage of *Callorhinchus* is not shown. Some elements of *Acanthodes*, *Pteronisculus* and *Eusthenopteron* are outlined in dashes or are uncoloured, because of their uncertain existence or uncertain homologies (Supplementary Notes). hb ant., hypobranchials anteriorly directed; hb post., hypobranchials posteriorly directed; ib ant., infrapharyngobranchial anteriorly directed; ib post., infrapharyngobranchial posteriorly directed. Light orange, compound infrapharyngobranchial 4, 5 (*Scyliorhinus*) or 3, 4, 5 (*Callorhinchus*) plus epihyal 5 (*Scyliorhinus*) or 4, 5 (*Callorhinchus*). Other colours and abbreviations are the same as in Fig. 1. Shared derived characters of each clade are listed at the corresponding nodes. The layout is not intended to reflect strict chronological history. Not to scale. Daggers indicate fossil taxa.

present in modern basal actinopterygians (Polypteriformes, Lepisosteiformes, possibly Acipenseriformes)[20] and in the Triassic actinopterygian *Pteronisculus*[8]. It has been suggested that these accessory cartilages are serially homologous with the symplectic or the interhyal of the hyoid arch[8,20]. Their occurrence in a stem chondrichthyan considerably expands their distribution, suggesting that they are a gnathostome symplesiomorphy. No traces of extrabranchial cartilages[3] were found in *Ozarcus*.

Ozarcus considerably broadens our understanding of visceral arch morphology in jawed vertebrates. Several features can now be reinterpreted as primitive for crown gnathostomes (Fig. 3 and Supplementary Notes), including: presence of supra- and infrapharyngobranchials, anteriorly directed infrapharyngobranchials, chevron arrangement of branchial arches along the basibranchial copula, all hypobranchials directed anteriorly, accessory cartilages between the cerato- and epibranchials, and paired hypohyals. Additionally, several features are reformulated here as potential synapomorphies of crown chondrichthyans, including a single (infra)pharyngobranchial element per arch, infrapharyngobranchials directed posteriorly, hypobranchials of the branchial arches concentrated at the anterior part of the basibranchial copula, absence of accessory cartilages, and absence of paired hypohyals anterior to the ceratohyals. The presence of a basihyal and fusion of epibranchial 5 with pharyngobranchial 4 may be apomorphic for, or within, crown chondrichthyans (Supplementary Notes). Having most of the hypobranchials directed posteriorly is interpreted as a synapomorphy of living elasmobranchs and extinct hybodont sharks (collectively known as euselachians[29]), not shared with modern chimaeroids.

Ozarcus challenges a widely held opinion that the ancestral state of the gnathostome branchial arch resembled the Σ-shape arches of

modern sharks, which have posteriorly directed pharyngo- and hypobranchial elements. Instead, the <-shaped arrangement of the bony fishes, with anteriorly directed pharyngo- and hypobranchial elements, is likely to be primitive (Fig. 3). *Ozarcus* thus shows a novel combination of chondrichthyan and osteichthyan characters, thereby demonstrating that the most recent common ancestor of crown gnathostomes possessed an osteichthyan-like branchial apparatus. Our findings cast doubt on the traditional view of visceral arch evolution that modern chondrichthyans mirror the ancestral morphotype of jawed vertebrates. Bony fishes and stem chondrichthyans may have more to tell us about our first jawed ancestors than do living sharks.

METHODS SUMMARY

The *O. mapesae* specimens were scanned on the ID19 beamline of the European Synchrotron Radiation Facility (ESRF). The scan parameters were as follows: voxel size 30.3 μm; single propagation distance of 3 m; BM5 beam, filtered by 90 mm of aluminium and 0.1 mm of molybdenum; effective energy 106 KeV; attenuation protocol with beam profiler; CCD FreLon 2K14 detector camera; LuAG 750 μm cintillator; double scan, half acquisition, with 5,000 projections of 0.1 s, phase retrieval using a Paganin process. The final reconstruction (60 μm voxel size) was obtained after binning. Volumes were reconstructed using ESRF software PyHST. Segmentation and three-dimensional rendering were performed with MIMICS 15.01 64-bit software. Details of both the *S. canicula* and *Callorhinchus milii* specimens are provided elsewhere[30].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 December 2013; accepted 3 March 2014.
Published online 16 April 2014.

1. Gegenbaur, C. *Elements of Comparative Anatomy* (Macmillan, 1878).
2. Janvier, P. *Early Vertebrates* (Oxford Univ. Press, 1996).
3. Mallatt, J. Ventilation and the origin of jawed vertebrates: a new mouth. *Zool. J. Linn. Soc.* **117,** 329–404 (1996).
4. Kuratani, S. Evolution of the vertebrate jaw from developmental perspective. *Evol. Dev.* **14,** 76–92 (2012).
5. De. Beer, G. R. *The Development of the Vertebrate Skull* (Clarendon, 1937).
6. Nelson, G. J. in *Nobel Symposium 4, Current Problems of Lower Vertebrate Phylogeny* (ed. Ørvig, T.) 129–143 (Almqvist & Wiksell, 1968).
7. Nelson, G. J. Gill arches and the phylogeny of fishes, with notes on the classification of vertebrates. *Bull. Am. Mus. Nat. Hist.* **141,** 477–552 (1969).
8. Nielsen, E. *Studies on Triassic fishes from East Greenland. I.* Glaucolepis *and* Boreosomus Vol. 1 (C. A. Reitzel, 1942).
9. Gardiner, B. G. The relationships of the palaeoniscid fishes, a review based on new specimens of *Mimia* and *Moythomasia* from the Upper Devonian of Western Australia. *Bull. Br. Mus. Nat. Hist. Geol.* **37,** 173–428 (1984).
10. Miles, R. S. In *Interrelationships of Fishes* (eds Greenwood, P. H., Miles, R. S. & Patterson, C.) 63–103 (Academic, 1973).
11. Heidtke, U. H. J. & Krätschmer, K. *Gladbachus adentatus* nov. gen. et sp., a primitive shark from the Upper Givetian Upper Middle Devonian of Bergisch Gladbach - Paffrath Basin Rhenish Slate Mountains [in German]. *Mainzer Geowissenschaftliche Mitteilungen* **30,** 105–122 (2001).
12. Jarvik, E. *Basic Structure and Evolution of Vertebrates* Vol. 1 (Academic, 1980).
13. Grogan, E. D. & Lund, R. *Debeerius ellefseni* (fam. nov., gen. nov., spec. nov.), an autodiastylic chondrichthyan from the Mississippian Bear Gulch Limestone of Montana (USA), the relationships of the Chondrichthyes, and comments on gnathostome evolution. *J. Morphol.* **243,** 219–245 (2000).
14. Coates, M. I. & Sequeira, S. E. K. A new stethacanthid chondrichthyan from the Lower Carboniferous of Bearsden, Scotland. *J. Vertebr. Paleontol.* **21,** 438–459 (2001).
15. Zangerl, R. & Williams, M. E. New evidence of the nature of the jaw suspension in Palaeozoic anacanthus sharks. *Palaeontology* **18,** 333–341 (1975).
16. Watson, D. M. S. The acanthodian fishes. *Philos. T. Roy. Soc. B* **228,** 49–146 (1937).
17. Dick, J. R. On the Carboniferous shark *Tristychius arcuatus* Agassiz from Scotland. *T. Roy. Soc. Edin.* **70,** 63–108 (1978).
18. Patterson, C. Significance of fossils in determining evolutionary relationships. *Annu. Rev. Ecol. Syst.* **12,** 195–223 (1981).
19. Zhu, M. *et al.* A Silurian placoderm with osteichthyan-like marginal jaw bones. *Nature* **502,** 188–193 (2013).
20. Carvalho, M., Bockmann, F. A. & de Carvalho, M. R. Homology of the fifth epibranchial and accessory elements of the ceratobranchials among gnathostomes: insights from the development of Ostariophysans. *PLoS ONE* **8,** e62389 (2013).
21. Friedman, M. & Brazeau, M. D. A reappraisal of the origin and basal radiation of the Osteichthyes. *J. Vertebr. Paleontol.* **30,** 36–56 (2010).
22. Shirai, S. *Squalean Phylogeny: A New Framework of ''Squaloid'' Sharks and Related Taxa* (Hokkaido Univ. Press, 1992).
23. Pradel, A., Tafforeau, P., Maisey, J. G. & Janvier, P. A new Paleozoic Symmoriiformes (Chondrichthyes) from the Late Carboniferous of Kansas (USA) and cladistic analysis of early chondrichthyans. *PLoS ONE* **6,** e24938 (2011).
24. Mapes, R. H. *Carboniferous and Permain Bactritoidia (Cephalopoda) in North America* (Univ. of Kansas Paleontological Institute, 1979).
25. Mallatt, J. Shark pharyngeal muscle and early vertebrate evolution. *Acta Zool.* **78,** 279–294 (1997).
26. Holmgren, N. Studies on the head in fishes. Part III. The phylogeny of elasmobranch fishes. *Acta Zool.* **23,** 129–261 (1942).
27. Lane, J. A. & Maisey, J. G. The visceral skeleton and jaw suspension in the durophagous hybodontid shark *Tribodus limae* from the Lower Cretaceous of Brazil. *J. Vertebr. Paleontol.* **86,** 886–905 (2012).
28. Davis, S. P., Finarelli, J. A. & Coates, M. I. *Acanthodes* and shark-like conditions in the last common ancestor of modern gnathostomes. *Nature* **486,** 247–250 (2012).
29. Maisey, J. G. What is an 'elasmobranch'? The impact of palaeontology in understanding elasmobranch phylogeny and evolution. *J. Fish Biol.* **80,** 918–951 (2012).
30. Pradel, A., Didier, D. A., Casane, D., Tafforeau, P. & Maisey, J. G. Holocephalan embryo provides new information on the evolution of the glossopharyngeal nerve, metotic fissure and parachordal plate in gnathostomes. *PLoS ONE* **8,** e66988 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank staff at the ID19 beamline at the ESRF for assistance, and F. Ippolito (AMNH) for photographs of the specimens. The main work was supported by the H. R. & E. Axelrod Research Chair in paleoichthyology at the AMNH.

Author Contributions J.G.M. and A.P. conceived the project. A.P. performed computerized microtomography restorations. A.P., J.G.M. and J.M. interpreted the results and prepared the manuscript. P.T. performed synchrotron computerized microtomography on the material. R.H.M. did the fieldwork.

Author Information Data have been deposited in ZooBank under the following LSIDs: urn:lsid:zoobank.org:pub:D80955F3-7C21-457C-BA68-9C70F466E913 (article); urn:lsid:zoobank.org:act:F7EDD4EB-6700-4583-8569-9BAC89B3A945 (genus); and urn:lsid:zoobank.org:act:6ACCB5D2-E0D7-4BFD-85D6-9C8250515614 (species). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.P. (apradel@amnh.org) or J.G.M. (maisey@amnh.org).

LETTER

doi:10.1038/nature13377

Bacterial phylogeny structures soil resistomes across habitats

Kevin J. Forsberg[1]*, Sanket Patel[1,2]*, Molly K. Gibson[1], Christian L. Lauber[3], Rob Knight[4,5], Noah Fierer[3,6] & Gautam Dantas[1,2,7]

Ancient and diverse antibiotic resistance genes (ARGs) have previously been identified from soil[1–3], including genes identical to those in human pathogens[4]. Despite the apparent overlap between soil and clinical resistomes[4–6], factors influencing ARG composition in soil and their movement between genomes and habitats remain largely unknown[3]. General metagenome functions often correlate with the underlying structure of bacterial communities[7–12]. However, ARGs are proposed to be highly mobile[4,5,13], prompting speculation that resistomes may not correlate with phylogenetic signatures or ecological divisions[13,14]. To investigate these relationships, we performed functional metagenomic selections for resistance to 18 antibiotics from 18 agricultural and grassland soils. The 2,895 ARGs we discovered were mostly new, and represent all major resistance mechanisms[15]. We demonstrate that distinct soil types harbour distinct resistomes, and that the addition of nitrogen fertilizer strongly influenced soil ARG content. Resistome composition also correlated with microbial phylogenetic and taxonomic structure, both across and within soil types. Consistent with this strong correlation, mobility elements (genes responsible for horizontal gene transfer between bacteria such as transposases and integrases) syntenic with ARGs were rare in soil by comparison with sequenced pathogens, suggesting that ARGs may not transfer between soil bacteria as readily as is observed between human pathogens. Together, our results indicate that bacterial community composition is the primary determinant of soil ARG content, challenging previous hypotheses that horizontal gene transfer effectively decouples resistomes from phylogeny[13,14].

Functional metagenomic selections permit deep interrogation of resistomes and can identify full-length, functionally verified ARGs without requiring sequence similarity to previously identified genes[2–4,16]. We constructed metagenomic libraries averaging 13.8 ± 8.3 (mean ± s.d.) gigabases (Gb) by shotgun cloning DNA fragments 1–5 kilobases (kb) long from 18 soils (Supplementary Table 1) into *Escherichia coli*, and screened these libraries for resistance against 18 antibiotics representing 8 drug classes. Resistance was conferred against 15 of the 18 antibiotics tested (Extended Data Fig. 1, Supplementary Table 2), and DNA fragments conferring resistance were sequenced, assembled, and annotated with PARFuMS[4] (see Methods).

We assembled 4,655 contigs over 500 base pairs (bp) in length (Fig. 1a, N50 size = 2.25 kb, or average contig length >1.76 kb) containing 8,882 open reading frames (ORFs) larger than 350 bp (Supplementary Data 1). Using profile Hidden Markov Models (HMMs), we annotated 2,895 of these 8,882 ORFs as ARGs (see Methods). Underscoring the immense functional diversity of soil resistomes, the identified soil ARGs were largely dissimilar from ARGs in public repositories (Fig. 1b). Only 15 soil ARGs (0.5%) have perfect amino acid identity to entries in the NCBI protein database, with just three having >99% nucleotide identity to nucleotide sequences in NCBI. In contrast, the average amino acid identity of all ARGs to their closest homologue in NCBI is only 61.1 ± 15.3%. Although we recently described cultured soil bacteria harbouring ARGs with perfect nucleotide identity to those in human pathogens[4], this phenomenon appears to be the exception rather than the rule: only one soil ARG from our current data set shares perfect nucleotide identity with a pathogen (NCBI accession number AY664504).

Our sampling depth (Extended Data Fig. 1) surpasses previous functional interrogations of soil metagenomes[8–10,16,17], permitting an unparalleled comparison of ARGs across soil types. Emphasizing the diversity recovered by our functional selections, 29% of assembled contigs over 1,500 bp did not contain an ORF that could be confidently assigned a known resistance function, representing a large repertoire of potentially novel ARGs. The ORFs assigned to known ARG functions represent all classical mechanisms of antibiotic resistance (Fig. 1c): antibiotic efflux, antibiotic inactivation and target protection/redundancy[15].

Resistance to amphenicol and tetracycline antibiotics occurred predominantly via the action of drug transporters, of which the majority belonged to the major facilitator superfamily (Fig. 1c). In contrast, selections with aminoglycoside and β-lactam antibiotics most frequently uncovered ARGs with antibiotic-inactivating capabilities, via covalent modification of aminoglycosides and enzymatic degradation of β-lactams (Fig. 1c). Excluding selections with trimethoprim and D-cycloserine (for which the ARGs selected were predominantly overexpressed target alleles from diverse bacterial lineages[16], Extended Data Fig. 2), β-lactamases were the most frequently encountered soil ARGs, mirroring observations from hospital settings[18]. We observed metallo-β-lactamases most frequently, followed by Ambler class-A and class-D enzymes (Extended Data Fig. 3).

We predicted the source phylum of each functionally selected resistance contig greater than 500 bp using a composition-based, semi-supervised, taxonomic binning algorithm[19]. Proteobacteria and Actinobacteria were the most prevalent predicted phyla, and each contained ARG families of all major resistance mechanisms[15] (Fig. 2a). Major facilitator superfamily transporters and β-lactamases showed the strongest, and most orthogonal, relationships with predicted bacterial phyla (Fig. 2b, Extended Data Fig. 4, Supplementary Table 3). β-lactamases were enriched within Verrucomicrobia, Acidobacteria and Cyanobacteria contigs, while major facilitator superfamily transporters were largely absent from Acidobacteria and were enriched among Actinobacteria and Proteobacteria contigs (Supplementary Table 3).

To quantitatively compare soil resistomes at higher resolution, a count matrix of unique gene sequences per functional annotation (that is, ARG family) was generated by summing across all antibiotic selections per soil and normalizing these counts to metagenomic library size (see Methods). The number of unique ARGs was significantly higher ($P < 0.01$, Wilcoxon rank sum test) in Cedar Creek (CC) grassland soils compared to agricultural soils from Kellogg Biological Station (KBS). Our selections resolved functional differences between CC and KBS soils regardless of whether Bray–Curtis distances were calculated using only ARG families (Fig. 3a) or all captured gene functions (Extended Data Fig. 5). Major facilitator superfamily transporters and β-lactamases were higher

[1]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St Louis, Missouri 63108, USA. [2]Department of Pathology and Immunology, Washington University School of Medicine, St Louis, Missouri 63110, USA. [3]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado 80309, USA. [4]Department of Chemistry and Biochemistry and BioFrontiers Institute, University of Colorado, Boulder, Colorado 80309, USA. [5]Howard Hughes Medical Institute, Boulder, Colorado 80309, USA. [6]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado 80309, USA. [7]Department of Biomedical Engineering, Washington University, St Louis, Missouri 63130, USA.
*These authors contributed equally to this work.

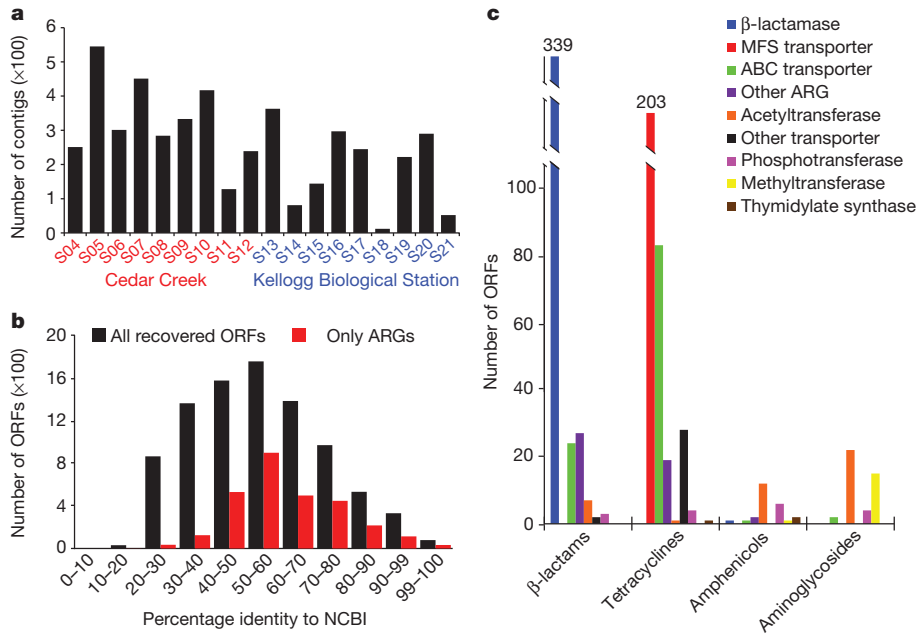


Figure 1 | Functional selections of 18 soil libraries yield diverse ARGs. **a**, Bar chart depicting contigs >500 bp across all antibiotic selections from CC (red) and KBS (blue) libraries. **b**, Amino acid identity between all ORFs (black, $n = 8{,}882$) or ARGs only (red, $n = 2{,}895$) and their top hit in the NCBI protein database. **c**, Total ARGs by antibiotic class (ARG types in the key); the *y* axis shows number of ORFs. MFS, major facilitator superfamily; ABC, ATP-binding cassette.

at CC compared to KBS, and ARG families of these resistance mechanisms most significantly differed between these soils (Supplementary Table 4). Only 2.6% of ARGs were shared across at least two soils at ≥99% nucleotide identity (Supplementary Table 5), with significantly more inter-soil sharing at CC versus KBS ($P < 0.05$, Fisher's exact test). No ARGs were shared between CC and KBS soils (≥99% nucleotide identity), and only two ARG mechanisms were observed in every soil (β-lactamase, major facilitator superfamily transporter).

We sampled CC soils across a gradient of added nitrogen (N) fertilizer. Similar to phylogenetic differences observed in community composition across the N gradient[20], we found that ARG composition of soils receiving higher N levels differed from the composition observed in other CC soils (Fig. 3b). These differences do not arise from a change in the number of unique ARGs between high-N and other soils ($P = 0.9$, Wilcoxon rank sum test), but rather were due to differences in relative proportions of ARG families in these soils (Supplementary Table 6). In high-N soils, β-lactamases were depleted whereas membrane transporters were enriched (Supplementary Table 6). High N levels favour particular organisms (for example, copiotrophs[8,10,20]), causing shifts in bacterial abundances, which in turn probably affect resistome composition.

We calculated differences in community structure of these CC and KBS soils using 16S ribosomal RNA gene sequences[8] (Extended Data Fig. 6). All bacterial phyla that were abundant (>3% relative abundance) in the samples, as determined by 16S rRNA gene sequencing, were also well-represented (>4% relative abundance) among phyla inferred from resistance-conferring contigs (Extended Data Fig. 7). Actinobacteria (which are characterized by GC-rich genomes and produce antibiotics in the soil[21]) were most enriched in resistance-conferring contigs relative

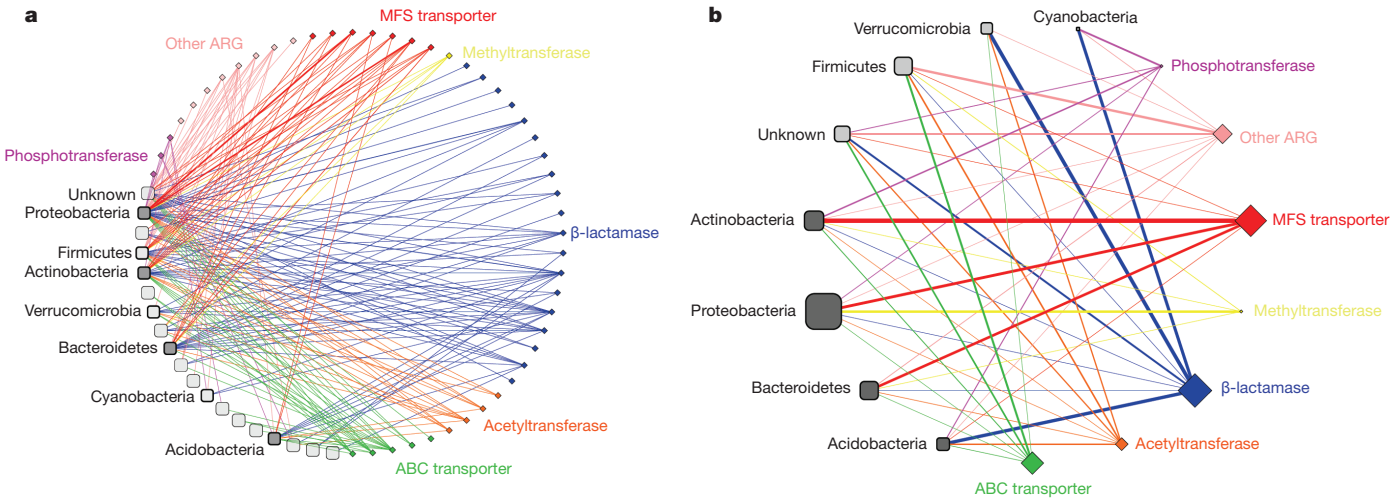


Figure 2 | Resistance is encoded by diverse soil phyla. **a**, Network of predicted bacterial phyla for each ARG used in cross-soil comparisons ($n = 880$). Edge thickness indicates number of ARGs within an ARG family (diamonds) from a predicted phylum (rounded squares). Phyla containing >15 ARGs are labelled, and are shaded dark grey at >3% 16S rRNA abundance. **b**, Simplified network of general ARG mechanisms; edge thickness represents significance of phylum and ARG mechanism co-occurrence (Fisher's exact test, line width increases with ranked significance). Node size indicates number of ARGs (diamonds) or contig count (rounded squares).

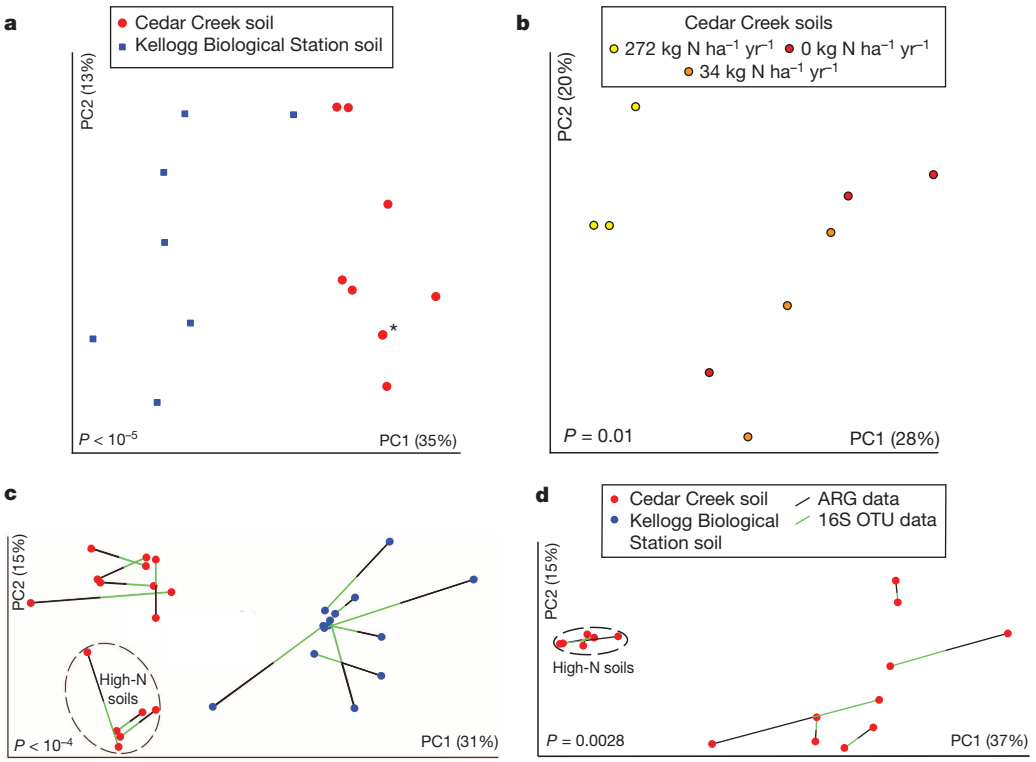


Figure 3 | Resistomes correlate with phylogeny across soil type and nitrogen amendment. **a**, **b**, Principal coordinate analysis (PCoA) plots depict Bray–Curtis distances between soils, using unique ARG counts. **a**, Resistomes from CC (red) and KBS (blue) soils cluster separately ($P < 10^{-5}$, analysis of similarity, ANOSIM). An asterisk denotes two soils with near-identical coordinates. **b**, CC soils amended with high N-levels cluster separately from other CC soils ($P = 0.01$, ANOSIM). **c**, **d**, Procrustes analyses depict significant correlation between ARG content (Bray–Curtis) and bacterial composition (Bray–Curtis) for CC (red) and KBS (blue) soils (**c**) and only CC soils (**d**). OTU, operational taxonomic unit.

to 16S rRNA gene abundance, whereas levels of Proteobacteria were similar in both data sets (Extended Data Fig. 7). Thus, any phylogenetic bias in functional selections due to heterologous expression in *E. coli* (a member of Proteobacteria) is minimal compared to the effect of the ARG content of source bacteria.

We next tested for correlations between soil resistomes and community composition. When both CC and KBS soils were considered, Bray–Curtis distances calculated from normalized ARG counts significantly correlated with bacterial operational taxonomic units inferred from 16S rRNA sequence data, whether taxonomic (Bray–Curtis) or phylogenetic (weighted and unweighted) dissimilarity metrics were used (Mantel tests, $P < 0.05$, Supplementary Table 7). Visualized by Procrustes analyses, both the ARG content and bacterial composition of CC and KBS soils clustered by sampling site, consistently displaying highly significant goodness-of-fit measures (Fig. 3c, Extended Data Fig. 8, Supplementary Table 8). Within sampling sites, the variability in phylogenetic community composition differed (Supplementary Table 9): more diversity was observed across CC soils than in KBS soils (Extended Data Fig. 6). Because of this disparity, we observed a significant within-site correlation between ARG content and community composition in CC soils (Supplementary Tables 7 and 8; Fig. 3d, Extended Data Fig. 8), but not in KBS soils (Extended Data Fig. 9).

The strong correlation between soil ARG content and bacterial composition suggests that the horizontal gene transfer (HGT) of ARGs is not sufficiently frequent to obscure their association with bacterial genomes. Corroborating this notion, soil ARGs show limited genetic potential for horizontal exchange. Only 0.42% of ORFs from our functional selections were predicted mobility elements (such as transposases, integrases and recombinases; Extended Data Fig. 10), and none of these genes were co-localized with an ARG containing >72% amino acid identity to a protein in NCBI. The limited mobility of the soil resistome may explain why ARGs are rarely shared between soil and human pathogens[4,22]. In contrast to soils, ARGs in pathogens often share near-perfect sequence identity[18], with origins that can be traced to the emergence of a single genotype disseminated broadly via HGT[23,24].

To test the hypothesis that ARGs in the soil have less potential for HGT than those in human pathogens, we compared ARGs from our functional selections to ARGs in fully sequenced genomes from 433 common human pathogens and 153 non-pathogenic soil organisms[13] (Supplementary Data 2). We modelled functional selections from each genome collection based on the fragment-size distribution observed in our soil selections (see Methods), and calculated the proportion of DNA fragments from each simulation that contained a predicted mobility element. Signatures of HGT were significantly more frequent in pathogen genomes than in soil genomes or soil selections (Fig. 4a). Importantly, we detected no difference in the HGT potential of ARGs between the two soil data sets (Fig. 4a), supporting the generality of the conclusions drawn from our soil functional selections. As the genetic distance from an ARG increased, the incidence of mobility elements in pathogen genomes was always higher than in soil genomes or functionally selected metagenomes (Fig. 4b), indicating that the higher potential for HGT seen in pathogens is independent of DNA fragment size or the method by which soil resistance is interrogated. Interestingly, enriching for multidrug-resistant Proteobacteria in the soil[4,25] (which are frequently encountered as opportunistic pathogens in hospital settings[26]) increases the detection of shared resistance between soil and clinic[4], suggesting that they may represent a major conduit through which ARGs move between these environments.

Unlike most hospital settings, soils contain a huge diversity of ARGs[1–3,16,22], and therefore increasing antibiotic exposure (as has occurred over the past 70 years[27]) may favour pre-existing genotypes[1] rather than the acquisition of new ARGs[4]. This key distinction explains our observation that, despite extensive sampling, very little evidence exists for HGT of ARGs across soil communities. Indeed, our evidence points to phylogeny, rather

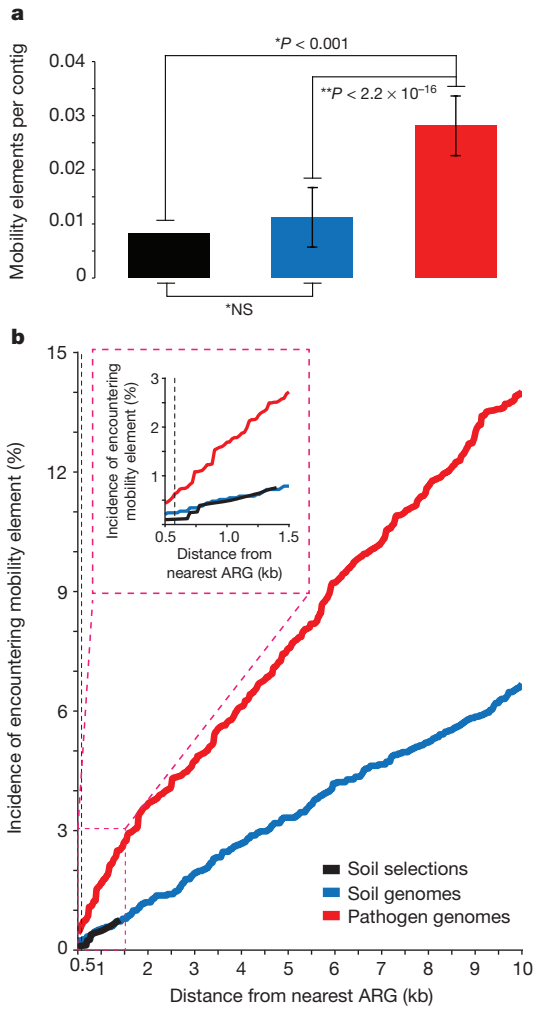


Figure 4 | Pathogen ARGs show higher HGT potential than soil ARGs. **a**, Mobility elements syntenic with ARGs are proportionally higher in pathogens ($n = 433$) than soil genomes ($n = 153$) or soil functional selections. *Significance determined from 1,000 Monte Carlo simulations; **significance determined by Student's *t*-test. NS, not significant. Error bars depict two standard deviations from the mean. **b**, Pathogens show significantly increased HGT potential relative to soil genomes and soil selections at all distances (20-bp intervals) greater than 580 bp (dashed line, $P < 0.05$, Fisher's exact test). The inset depicts mobility elements encountered within 1.5 kb of ARGs, demonstrating that data from soil selections resemble soil genomes.

than HGT[13,14], as the primary determinant of soil resistome content. Therefore, as bacterial type and diversity change across soils[8–10,20], so too do their associated ARGs, resulting in resistomes that may respond to anthropogenic modulations (for example, nitrogen fertilizer) that do not possess obvious antibiotic-related properties.

METHODS SUMMARY

Metagenomic DNA was extracted from all soils using the PowerMax soil DNA isolation kit (MoBio Laboratories). Small-insert metagenomic libraries were created by shearing this DNA into 1–5-kb fragments before ligation into the pZE21 expression vector[28] and electroporation into *E. coli* MegaX cells (Invitrogen). These libraries were plated on Mueller–Hinton agar containing one of 18 antibiotics, grown overnight at 37 °C, and resistant colonies collected in a liquid cell slurry. Metagenomic fragments conferring resistance were amplified via polymerase chain reaction (PCR) using this slurry as template, and barcoded libraries prepared from these amplicons for 101-bp paired-end sequencing using the Illumina HiSeq2000. Paired-end reads were assembled into resistance-conferring fragments using PARFuMs[4] and annotated using profile HMM databases. Cross-soil resistome comparisons were performed using Bray–Curtis distances calculated from a count matrix of unique gene sequences per ARG family, generated by summing across all antibiotic selections for a given soil and normalizing these counts to metagenomic library size. Taxonomies of assembled fragments were assigned using RAIphy[19], while the 16S rRNA gene sequence data generated previously were processed as described[8]. Genomes of human pathogens and non-pathogenic soil bacteria were collected based on published environmental metadata[13]. Functional selections were modelled from each genome collection based on the fragment-size distribution observed in our soil selections.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 February; accepted 14 April 2014.
Published online 21 May 2014.

1. D'Costa, V. M. *et al.* Antibiotic resistance is ancient. *Nature* **477,** 457–461 (2011).
2. Allen, H. K., Moe, L. A., Rodbumrer, J., Gaarder, A. & Handelsman, J. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J.* **3,** 243–251 (2009).
3. Allen, H. K. *et al.* Call of the wild: antibiotic resistance genes in natural environments. *Nature Rev. Microbiol.* **8,** 251–259 (2010).
4. Forsberg, K. J. *et al.* The shared antibiotic resistome of soil bacteria and human pathogens. *Science* **337,** 1107–1111 (2012).
5. Wright, G. D. Antibiotic resistance in the environment: a link to the clinic? *Curr. Opin. Microbiol.* **13,** 589–594 (2010).
6. Benveniste, R. & Davies, J. Aminoglycoside antibiotic-inactivating enzymes in actinomycetes similar to those present in clinical isolates of antibiotic-resistant bacteria. *Proc. Natl Acad. Sci. USA* **70,** 2276–2280 (1973).
7. Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnol.* **31,** 814–821 (2013).
8. Fierer, N. *et al.* Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J.* **6,** 1007–1017 (2012).
9. Fierer, N. *et al.* Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl Acad. Sci. USA* **109,** 21390–21395 (2012).
10. Fierer, N. *et al.* Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. *Science* **342,** 621–624 (2013).
11. Muegge, B. D. *et al.* Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332,** 970–974 (2011).
12. Zaneveld, J. R., Lozupone, C., Gordon, J. I. & Knight, R. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res.* **38,** 3869–3879 (2010).
13. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480,** 241–244 (2011).
14. Stokes, H. W. & Gillings, M. R. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS Microbiol. Rev.* **35,** 790–819 (2011).
15. Walsh, C. Molecular mechanisms that confer antibacterial drug resistance. *Nature* **406,** 775–781 (2000).
16. Pehrsson, E. C., Forsberg, K. J., Gibson, M. K., Ahmadi, S. & Dantas, G. Novel resistance functions uncovered using functional metagenomic investigations of resistance reservoirs. *Front. Microbiol.* **4,** 145 (2013).
17. Delmont, T. O. *et al.* Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J.* **6,** 1677–1687 (2012).
18. Jacoby, G. A. & Munoz-Price, L. S. The new beta-lactamases. *N. Engl. J. Med.* **352,** 380–391 (2005).
19. Nalbantoglu, O. U., Way, S. F., Hinrichs, S. H. & Sayood, K. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinform.* **12,** 41 (2011).
20. Ramirez, K. S., Lauber, C. L., Knight, R., Bradford, M. A. & Fierer, N. Consistent effects of nitrogen fertilization on soil bacterial communities in contrasting systems. *Ecology* **91,** 3463–3470; discussion 3503–3414 (2010).
21. Ventura, M. *et al.* Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol. Mol. Biol. Rev.* **71,** 495–548 (2007).
22. Aminov, R. I. & Mackie, R. I. Evolution and ecology of antibiotic resistance genes. *FEMS Microbiol. Lett.* **271,** 147–161 (2007).
23. Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74,** 417–433 (2010).
24. Medeiros, A. A. Evolution and dissemination of beta-lactamases accelerated by generations of beta-lactam antibiotics. *Clin. Inf. Diseases* **24** (Suppl. 1), 19–45 (1997).
25. Dantas, G., Sommer, M. O., Oluwasegun, R. D. & Church, G. M. Bacteria subsisting on antibiotics. *Science* **320,** 100–103 (2008).
26. Boucher, H. W. *et al.* Bad bugs, no drugs: no ESKAPE! *Clin. Inf. Diseases* **48,** 1–12 (2009).
27. Knapp, C. W., Dolfing, J., Ehlert, P. A. & Graham, D. W. Evidence of increasing antibiotic resistance gene abundances in archived soils since 1940. *Environ. Sci. Technol.* **44,** 580–587 (2010).
28. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1–I2 regulatory elements. *Nucleic Acids Res.* **25,** 1203–1210 (1997).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Pesesky for access to and assistance with the data set used to benchmark RAlphy's performance, B. Wang for suggested improvements to Illumina library preparation, the Genome Technology Access Center at Washington University in St Louis for generating Illumina sequence data, M. Sherman for discussions on modelling pathogen HGT potential, and members of the Dantas laboratory for discussions on the results and analyses presented here. This work was supported by awards to G.D. through the Children's Discovery Institute (MD-II-2011-117), the International Center for Advanced Renewable Energy and Sustainability at Washington University, the National Academies Keck Futures Initiatives (Synthetic Biology, SB2), and the NIH Director's New Innovator Award (DP2-DK-098089). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. M.K.G. is supported by a Mr and Mrs Spencer T. Olin Fellowship for Women in Graduate Study at Washington University. K.J.F. received support from the NIGMS Cell and Molecular Biology Training Grant (GM 007067) and from the NHGRI Genome Analysis Training Program (T32 HG000045). K.J.F. and M.K.G. are NSF graduate research fellows (award number DGE-1143954).

Author Contributions N.F., C.L.L. and R.K. provided soils and 16S rRNA gene sequencing data; G.D. conceived the functional selections; S.P. created metagenomic libraries, performed functional selections, and prepared sequencing libraries; K.J.F. assembled sequence data from functional selections and annotated ARGs with assistance from M.K.G.; M.K.G. built the custom ARG profile HMM database; K.J.F. performed genomic and ecological analyses and wrote the manuscript with contributions from M.K.G., R.K., N.F. and G.D.

Author Information All assembled sequences have been deposited to Genbank with accession numbers KJ691878–KJ696532 and raw reads to SRA under the accession number SRP041174. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.D. (dantas@wustl.edu).

LETTER

doi:10.1038/nature13250

Epidermal Merkel cells are mechanosensory cells that tune mammalian touch receptors

Srdjan Maksimovic[1]*, Masashi Nakatani[1,2]*, Yoshichika Baba[1]*, Aislyn M. Nelson[1,3], Kara L. Marshall[1], Scott A. Wellnitz[3], Pervez Firozi[3], Seung-Hyun Woo[4], Sanjeev Ranade[4], Ardem Patapoutian[4] & Ellen A. Lumpkin[1,5,6]

Touch submodalities, such as flutter and pressure, are mediated by somatosensory afferents whose terminal specializations extract tactile features and encode them as action potential trains with unique activity patterns[1]. Whether non-neuronal cells tune touch receptors through active or passive mechanisms is debated. Terminal specializations are thought to function as passive mechanical filters analogous to the cochlea's basilar membrane, which deconstructs complex sounds into tones that are transduced by mechanosensory hair cells. The model that cutaneous specializations are merely passive has been recently challenged because epidermal cells express sensory ion channels and neurotransmitters[2,3]; however, direct evidence that epidermal cells excite tactile afferents is lacking. Epidermal Merkel cells display features of sensory receptor cells[4,5] and make 'synapse-like' contacts[5,6] with slowly adapting type I (SAI) afferents[7–9]. These complexes, which encode spatial features such as edges and texture[1], localize to skin regions with high tactile acuity, including whisker follicles, fingertips and touch domes. Here we show that Merkel cells actively participate in touch reception in mice. Merkel cells display fast, touch-evoked mechanotransduction currents. Optogenetic approaches in intact skin show that Merkel cells are both necessary and sufficient for sustained action-potential firing in tactile afferents. Recordings from touch-dome afferents lacking Merkel cells demonstrate that Merkel cells confer high-frequency responses to dynamic stimuli and enable sustained firing. These data are the first, to our knowledge, to directly demonstrate a functional, excitatory connection between epidermal cells and sensory neurons. Together, these findings indicate that Merkel cells actively tune mechanosensory responses to facilitate high spatio-temporal acuity. Moreover, our results indicate a division of labour in the Merkel cell–neurite complex: Merkel cells signal static stimuli, such as pressure, whereas sensory afferents transduce dynamic stimuli, such as moving gratings. Thus, the Merkel cell–neurite complex is an unique sensory structure composed of two different receptor cell types specialized for distinct elements of discriminative touch.

We first asked whether Merkel cells display touch-activated currents. GFP-expressing Merkel cells were dissociated from mouse epidermis and purified by flow cytometry for whole-cell recordings[10] (Fig. 1a, b). Merkel cells showed displacement-dependent inward currents (Fig. 1c, d), whereas keratinocytes lacked mechanosensitive currents over the same stimulus range ($n = 4$). Merkel-cell currents averaged 370 ± 80 pA at peak and 20 ± 6 pA at steady state (Extended Data Fig. 1 and Table 1). They displayed steep displacement operating ranges (1.6 ± 0.1 μm; Fig. 1d, e) and millisecond rise times (Fig. 1f and Extended Data Table 1), which were limited by stimulus-probe velocity. This suggests fast gating mechanisms, similar to force-gated channels in hair cells and invertebrate mechanosensory neurons[11].

We next asked whether biophysical properties of touch-activated currents matched those of identified mechanosensitive channels. Currents showed fast inactivation time constants (8 ± 2 ms; Fig. 1g), similar to those of dorsal root ganglia (DRG) neurons[12]. Mechanotransduction currents showed a linear current–voltage relation (reversal potential, 8 ± 3 mV, Fig. 1h), indicating non-rectifying, non-selective cation channels. Moreover, they were attenuated by Ruthenium red (RR), an antagonist of Piezo and TRP cation channels (Extended Data Fig. 1). These biophysical properties are consistent with mechanosensitive channels encoded by *Piezo* genes (Piezo2: $\tau_{inactivation} = 7 \pm 1$ ms; $E_{rev} = 9 \pm 2$ mV; RR block ~80%)[13]. Indeed, quantitative reverse-transcription polymerase chain reaction (qRT-PCR) analysis showed that Merkel cells express *Piezo1* and *Piezo2*, and that *Piezo2* is enriched in Merkel cells compared with total epidermal cells, which predominantly comprise keratinocytes (Fig. 1i).

Merkel cells also exhibited robust touch-evoked increases in cytoplasmic Ca^{2+} (Extended Data Fig. 1). As these cells were not voltage clamped, calcium signals probably reflected calcium entry through mechanotransduction channels, opening of voltage-activated calcium channels and subsequent calcium-induced calcium release, as is the case for hypotonic-activated responses in Merkel cells[10]. Collectively, our findings demonstrate for the first time that Merkel cells are capable of transducing touch stimuli into excitatory responses in the absence of sensory neurons or keratinocytes.

How might the Merkel cell's rapidly inactivating mechanotransduction currents lead to slowly adapting responses *in vivo*? Like hair cells and *Drosophila melanogaster* bristles[14], Merkel-cell mechanotransduction channels display steady-state currents that are ~10% of peak responses (Extended Data Fig. 1). These currents are likely to be amplified by voltage-activated calcium channels[4,10]. Indeed, an accompanying manuscript demonstrates that inward currents of ≤ 20 pA are sufficient to depolarize Merkel cells to voltage-activated ion-channel thresholds[15]. Moreover, computational modelling predicts that a rapidly adapting transduction current with a small steady-state component can account for SAI firing patterns[16].

We next tested whether activating Merkel cells in the intact skin is sufficient to excite tactile afferents. We used optogenetics to selectively depolarize Merkel cells without directly stimulating their associated sensory afferents (Fig. 2a). A previous microarray screen identified cholecystokinin (Cck) as a Merkel-cell-specific transcript in the epidermis[4]. To express Channelrhodopsin-2 (ChR2)[17] in Merkel cells *in vivo*, we crossed Cck-IRES-Cre mice[18] with mice harbouring a ChR2-tdTomato fusion at the *Gt(Rosa)26Sor* locus[19]. Heterozygote $Cck^{Cre/+}$;$ChR2^{loxP/+}$ mice showed strong expression of ChR2-tdTomato in touch-dome Merkel cells, whose fluorescence was easily identifiable in intact skin (Fig. 2b). Whole-cell recordings *in vitro* confirmed that $ChR2^{+}$ Merkel cells exhibited light-activated inward currents ($n = 5$, Extended Data Fig. 2). We confirmed the absence of ChR2 expression in SAI afferents via immunohistochemistry of skin cryosections (Fig. 2c) and whole-mounts (Extended Data Fig. 3a–e). ChR2 expression was not observed in any

[1]Department of Dermatology, Columbia University, New York, New York 10032, USA. [2]Graduate School of System Design and Management, Keio University, Yokohama 223-8526, Japan. [3]Department of Neuroscience, Baylor College of Medicine, Houston, Texas 77006, USA. [4]Howard Hughes Medical Institute, Molecular and Cellular Neuroscience, The Scripps Research Institute, La Jolla California 92037, USA. [5]Department of Physiology & Cellular Biophysics, Columbia University, New York, New York 10032, USA. [6]Program in Neurobiology & Behavior, Columbia University, New York, New York 10032, USA.
*These authors contributed equally to this work.

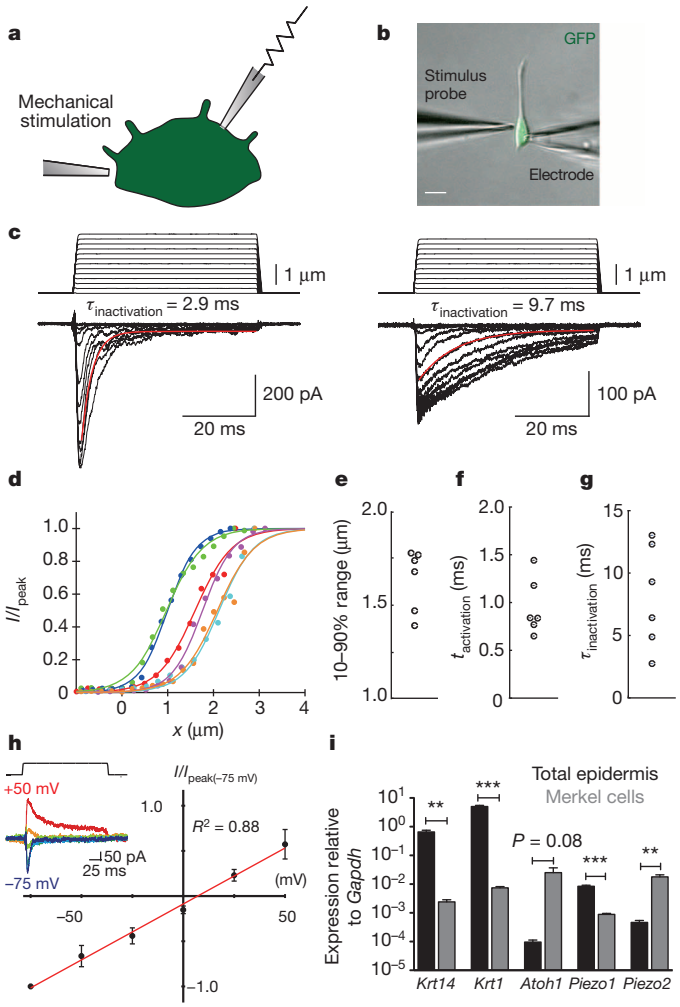


Figure 1 | Merkel cells exhibit touch-evoked ionic currents and preferentially express *Piezo2*. **a**, Schematic of whole-cell recordings from dissociated Merkel cells[10]. **b**, Merkel cells were identified based on nuclear GFP fluorescence and were stimulated by a glass probe driven by a piezoelectric actuator. Scale bar, 10 µm. **c**, Representative traces of mechanosensitive currents from two Merkel cells ($V_H = -70$ mV). Inactivation kinetics were estimated from single-exponential fits (red lines). **d**, Current–displacement relationships from individual Merkel cells ($n = 6$; indicated by distinct colours). Currents were normalized to peak and fitted with Boltzmann functions ($R^2 > 0.97$). **e**–**g**, Activation and inactivation kinetics and operating ranges for individual Merkel cells. **h**, Current–voltage relationship of mechanosensitive currents ($n = 5$). Peak current at each holding potential was normalized to peak current at −75 mV. Reversal potential was estimated by linear regression (red line; $R^2 = 0.88$). Inset: representative traces at each holding potential (denoted by colours). **i**, Expression (qRT-PCR) analysis of *Piezo1* and *Piezo2* transcripts in purified Merkel cells and total epidermal cell suspensions. Markers of keratinocytes (keratin 14 (*Krt14*) and keratin 1 (*Krt1*)) and Merkel cells (*Atoh1*) verified selective enrichment of Merkel cells. ***$P < 0.0001$, **$P \leq 0.001$ ($n = 4$ technical replicates).

cutaneous afferent ($n = 21$ mice). Epidermal ChR2-tdTomato expression was limited to Merkel cells; however, ChR2-tdTomato was observed in some dermal cell types (Extended Data Fig. 3).

To acquire touch-dome specific responses, we used laser-coupled fibre optics with a collimator lens to restrict illumination to touch domes (Fig. 2d). When touch domes were presented with 5-s light pulses, action potentials were observed in phase with light stimulation (Fig. 2e). All afferents that displayed light-evoked responses were touch-sensitive ($n = 12$), and were classified as SAI afferents based on physiological criteria. We confirmed that the touch- and light-evoked activity derived from single units with multidimensional spike-sorting[20] and by comparing spike waveforms (Fig. 2e, right). Light-evoked responses showed increased firing with increased light intensities (Fig. 2f). Illumination of tissue that lacked Merkel cells, including skin areas adjacent to touch domes and saphenous nerve trunks, did not evoke action potentials (Extended Data Fig. 4). Illumination of touch domes with green light (545 nm), which does not activate ChR2, also failed to elicit afferent firing. Light-evoked responses from SAI afferents ($n = 3$ units) were also observed when ChR2-tdTomato expression was driven by an epidermis-specific $K14^{Cre}$ transgene[21] (Extended Data Fig. 5), confirming that light-evoked responses require the presence of ChR2-expressing epidermal cells. Thus, we conclude that depolarization of epidermal Merkel cells is sufficient to excite action potentials in SAI afferents. To our knowledge, this is the first use of optogenetics in skin cells and the first functional proof of an excitatory connection between any epidermal cell type and tactile afferents in skin.

In SAI afferents, touch stimuli elicit biphasic responses with a dynamic phase characterized by high-frequency firing at touch onset, and a static phase characterized by sustained firing with highly variable inter-spike intervals (ISIs)[7–9]. To test whether Merkel-cell photostimulation recapitulated these properties, we recorded action potential trains elicited by 3-min light stimuli (Fig. 2g, h). In agreement with canonical SAI responses[7–9], light-evoked responses showed continuous firing throughout stimulation with coefficients of variation (CoV) of ISIs > 0.5 (mean ± s.d., 1.17 ± 0.14, $n = 12$). Interestingly, light-evoked responses lacked high-frequency firing at stimulus onset. Thus, selective activation of epidermal Merkel cells is sufficient to elicit action potential trains whose activity patterns mimic the static phase of touch-evoked SAI responses.

We next tested whether optogenetic silencing of Merkel cells inhibits touch-evoked firing in SAI afferents. We selectively expressed Archaerhodopsin-3 (ArchT), a green-light-sensitive, hyperpolarizing proton pump[22], in Merkel cells ($Cck^{Cre/+}$;$ArchT^{loxP/+}$; Extended Data Fig. 6). During 3-min displacements, touch domes were presented with a series of 10-s light pulses. We observed a tenfold reduction in median touch-evoked firing rates during light stimuli ($n = 3$; Fig. 2i, j). Inhibition grew progressively stronger during sustained displacement, becoming almost complete with successive light presentations (Fig. 2i). Thus, Merkel-cell depolarization is necessary for robust static phase firing in SAI afferents.

To determine whether Merkel cells also contribute to the dynamic phase of touch-evoked SAI responses, we analysed epidermal-specific *Atoh1* conditional knockout ($Atoh1^{CKO}$) mice ($K14^{Cre}$;$Atoh1^{lacZ/flox}$), which completely lack Merkel cells but have otherwise normal epidermis[23]. Touch domes were innervated by myelinated afferents that contain nodes of Ranvier, suggesting that they are capable of firing action potentials (Extended Data Fig. 7). A previous unbiased survey of touch-sensitive afferents reported a selective loss of SAI responses in $Hoxb1^{Cre}$;$Atoh1^{LacZ/flox}$ mice, which lack Merkel cells from development but retain innervation of touch domes and footpads[24]. Here, we used a targeted approach to analyse firing properties of afferents that selectively innervate touch domes, which were identified during recording based on FM1-43 fluorophore uptake[25]. We found that touch domes in $Atoh1^{CKO}$ mice were innervated by mechanosensitive Aβ afferents (conduction velocity: 10.2 – 18.5 m per s, $n = 6$); however, their firing patterns differed markedly from SAI responses[8,9] in control genotypes (Fig. 3a, b; $n = 5$). First, the static phase firing was truncated in $Atoh1^{CKO}$ mice compared with control afferents, which maintained firing throughout 5-s stimuli (Fig. 3a, b). Thus, responses of $Atoh1^{CKO}$ touch-dome afferents could be classified as intermediately adapting[9,26] (IA; Fig. 3c). Second, $Atoh1^{CKO}$ responses displayed markedly lower spike counts (Fig. 3d and Extended Data Table 2) and firing rates than control genotypes ($Atoh1^{CKO}$ mean ± s.d., 25 ± 8 Hz; control: 59 ± 25 Hz; $P = 0.004$). Notably, $Atoh1^{CKO}$ lacked high-frequency firing and short ISIs in both dynamic and static phases ($Atoh1^{CKO}$ peak firing rates, 79 ± 40 Hz; control, 238 ± 69 Hz; $P = 0.003$; Extended Data Fig. 8 and Extended Data Table 2). Together, these data indicate that touch-dome

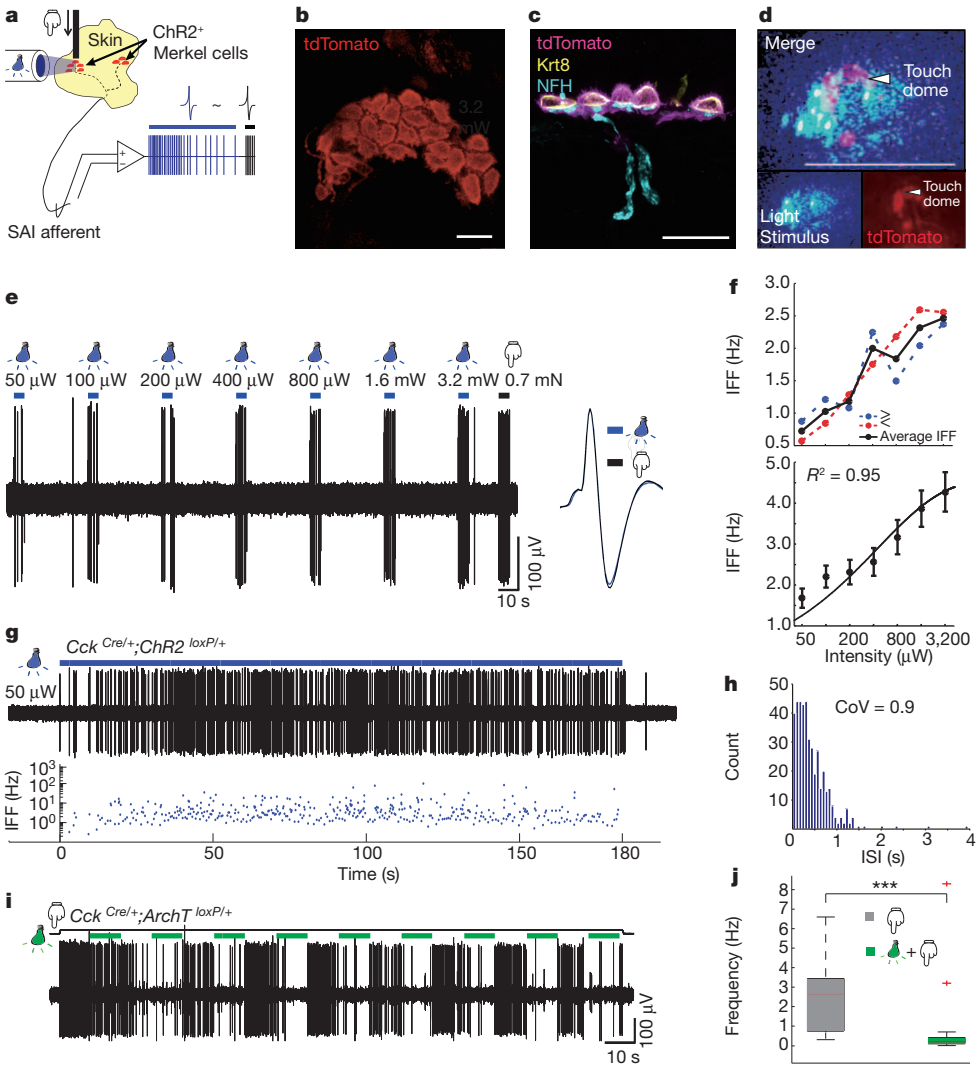


Figure 2 | Merkel cells are necessary and sufficient to elicit sustained action-potential trains in touch-dome afferents. **a**, Schematic of mouse *ex vivo* skin–nerve recordings. **b**, Confocal image of a ChR2-expressing touch dome in a living skin–nerve preparation. Scale bar, 20 µm. **c**, Immunostaining of skin cryosections shows expression of ChR2-tdTomato in Merkel cells (keratin 8, Krt8) but not in touch-dome afferents (Neurofilament heavy, NFH (also known as Nefh). Scale bar, 20 µm. **d**, During electrophysiological recording, ChR2-expressing Merkel cells and blue-light stimuli were imaged separately using different filter sets (bottom insets). Merged image illustrates the illuminated area (top panel). Confocal reconstruction of this touch dome is shown in Extended Data Fig. 2a–e. Scale bar, 200 µm. **e**, Light pulses of increasing intensities elicited phase-locked action potentials from the touch dome in **d** (left trace). Comparison of spike shapes evoked by light (blue) and touch (black) confirmed single-unit recording (right trace). **f**, Mean instantaneous firing frequency (IFF) versus light intensity for a single touch-dome afferent (top graph). Blue trace shows mean IFFs from **e**. Red trace shows mean IFFs evoked by light intensities presented in decreasing order. The averages of these stimuli (black trace) were analysed further in the lower part of the panel. Mean IFFs on a log-intensity scale ($n = 12$ single units; bottom graph). Data were fit with a four-parameter Weibull sigmoidal function ($R^2 = 0.95$). **g**, **h**, A sustained light-evoked response from the touch-dome afferent in **d** and **e** with corresponding ISI histogram. **i–j**, Optogenetic silencing of ArchT-expressing Merkel cells. **i**, Representative 3-min recording. **j**, Box-plot of firing rates during light-off ($n = 3$ units, $n = 20$ 10-s periods) and light-on (same units, $n = 23$ 10-s periods). Two outliers are firing rates from initial light-on periods. ***$P = 0.001$.

afferents have mechanosensory terminals capable of responding to touch, but do so with firing properties distinct from the canonical SAI response.

We next compared the responses of touch-dome afferents in *Atoh1*^{CKO} with epidermal-specific *Piezo2* knockout mice[15] (*Piezo2*^{CKO}). In the latter, Merkel cells and touch-dome afferents develop normally and are retained through adulthood[15]. Although these mutations disrupt distinct molecular pathways and cause different anatomical phenotypes, we found a remarkable degree of concordance between static-phase firing patterns. Mutant touch-dome afferents displayed a similar proportion of intermediately adapting responses (Fig. 3c)[15] and showed similar increases in mean ISIs during static displacement (Fig. 3e and Extended Data Fig. 8). Together, these data indicate that Piezo2-dependent Merkel-cell signalling is essential for proper SAI responses to sustained pressure.

Our study sheds new light on the role of Merkel cells in touch reception. Our findings demonstrate that Merkel cells are touch-sensitive cells that actively tune mechanosensory afferents by conferring two features of the SAI response: sustained responses and high-frequency firing. First, by maintaining firing throughout mechanical stimulation, slowly adapting afferents inform the brain about pressure[1,27]. Our optogenetic approach demonstrates that Merkel-cell activation elicits, and silencing reversibly suppresses, sustained SAI firing. This provides the first direct evidence that Merkel cells are not simply passive mechanical filters in the skin. Moreover, recordings from *Atoh1*^{CKO} and *Piezo2*^{CKO} mice show that SAI afferents cannot properly convey static phase information without intact Merkel cells. Second, during active tactile exploration, high-frequency firing is important for encoding object features (such as edges and curvature) with high information content[27,28]. Although

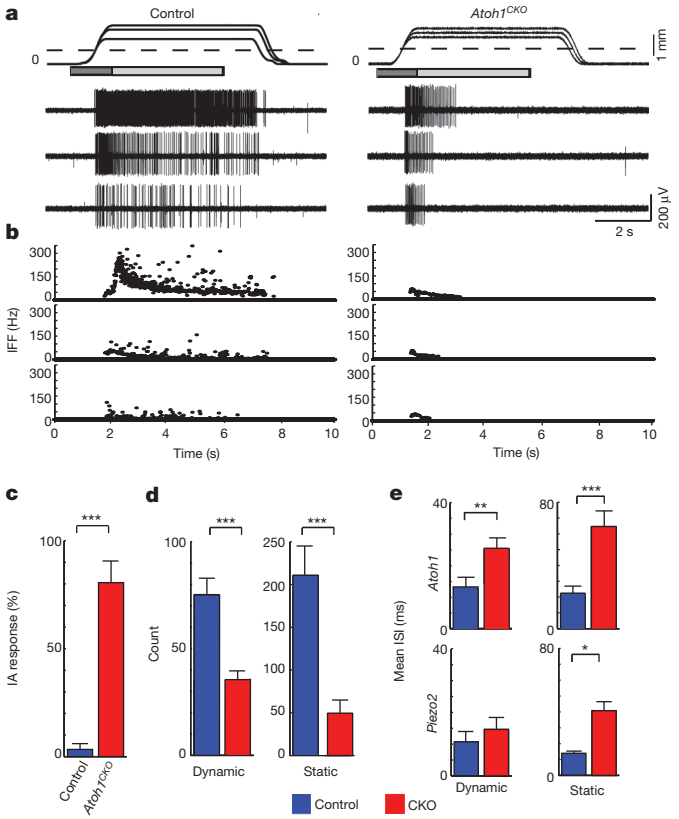


Figure 3 | *Atoh1*^{CKO} and *Piezo2*^{CKO} mice show intermediately adapting responses. **a**, Mechanically evoked responses from touch-dome afferents in control and *Atoh1*^{CKO} (*K14*^{Cre};*Atoh1*$^{lacZ/flox}$) mice. Top trace shows ramp-and-hold displacements at three magnitudes with corresponding action potential trains below. Dashed line marks the point of skin contact (0 mm). Boxes indicate dynamic (dark grey, 1.5 s after displacement command onset) and static phases for analysis (light grey, 4 s after the beginning of hold command). **b**, IFFs of the responses in **a**. **c**, Proportion of intermediately adapting (IA) responses to supra-threshold displacements in touch-dome afferents ($n = 5$ control and $n = 6$ *Atoh1*^{CKO} units). **d**, Mean number of spikes in dynamic and static phases of saturating responses. **e**, Mean interspike intervals (ISIs) of *Atoh1*^{CKO} ($n = 6$ units), *Piezo2*^{CKO} ($n = 6$ units) and their respective control genotypes ($n = 5$ units per genotype). Asterisks indicate statistically significant differences between mutant and control genotypes in Fisher's exact test (**c**) and Student's *t*-test (**d**, **e**). $*P < 0.05$; $**P < 0.02$; $***P < 0.01$.

touch-dome afferents in *Piezo2*^{CKO} showed similar dynamic responses to control mice, *Atoh1*^{CKO} afferents showed markedly reduced dynamic firing. Thus, Merkel cells perform *Piezo2*-independent functions that enhance dynamic firing, which is predicted to facilitate high spatio-temporal acuity of tactile perception. Consistent with this prediction, *Atoh1*^{CKO} mice display behavioural deficits in texture preference[29]. These effects of Merkel-cell loss could be because of differences in SAI afferent development or touch-dome mechanics, or *Piezo2*-independent signalling mechanisms.

Our findings support two models for how Merkel cells contribute to the SAI afferent's unique firing patterns. First, our data directly demonstrate that the Merkel cell–neurite complex is a compound sensory system with two receptor cell types that mediate different aspects of touch transduction[30] (Fig. 4). A similar division is found in the visual system, which also provides information on object shape and movement. In the retina, rods are optimized for low-light conditions and cones for high-acuity, colour vision. Second, the contribution of Merkel cells to dynamic firing indicates a previously unsuspected role in signal amplification, analogous to outer hair cells in the mammalian cochlea. These mechanosensory cells actively expend energy to tune the cochlea's frequency selectivity and mechanical sensitivity. A key question that

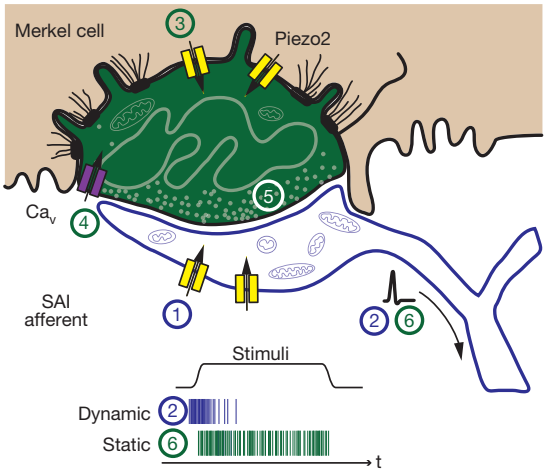


Figure 4 | Model of active Merkel-cell inputs in touch reception. Deformation of the skin opens mechanotransduction channels in SAI afferents (1) to initiate action potential firing at the onset of dynamic stimuli (2). The presence of Merkel cells boosts dynamic firing through Piezo2-independent mechanisms. Skin deformation simultaneously activates Piezo2-dependent[15] mechanotransduction channels in Merkel cells (3) to depolarize these cells, which produces calcium entry through voltage-activated calcium channels (Ca_V) (4) and release of unidentified neurotransmitters (5) that trigger sustained firing (6). Schematic modified from Iggo and Muir[7] and reproduced with permission of John Wiley & Sons.

remains is the nature of the excitatory mechanisms that convey signals between Merkel cells and SAI afferents.

METHODS SUMMARY

Experimental procedures were performed in compliance with the Institutional Animal Care and Use Committees of Columbia University and Baylor College of Medicine. See the full Methods for details of immunohistochemistry, electrophysiology, live-cell imaging and statistical analyses. Unless noted, statistical comparisons were performed with Student's *t*-test (unpaired, two-tailed), data are expressed as mean ± standard error of the mean (s.e.m.), and error bars denote s.e.m. Touch-sensitive units were classified as SAI afferents based on the following physiological criteria: low mechanical threshold (<1 mN), receptive field restricted to touch domes, sustained responses with irregular ISIs during ramp-and-hold stimuli, and Aβ conduction velocity.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 November 2013; accepted 13 March 2014.
Published online 6 April 2014.

1. Johnson, K. O. The roles and functions of cutaneous mechanoreceptors. *Curr. Opin. Neurobiol.* **11,** 455–461 (2001).
2. Lumpkin, E. A. & Caterina, M. J. Mechanisms of sensory transduction in the skin. *Nature* **445,** 858–865 (2007).
3. Kwan, K. Y., Glazer, J. M., Corey, D. P., Rice, F. L. & Stucky, C. L. TRPA1 modulates mechanotransduction in cutaneous sensory neurons. *J. Neurosci.* **29,** 4808–4819 (2009).
4. Haeberle, H. *et al.* Molecular profiling reveals synaptic release machinery in Merkel cells. *Proc. Natl Acad. Sci. USA* **101,** 14503–14508 (2004).
5. Maksimovic, S., Baba, Y. & Lumpkin, E. A. Neurotransmitters and synaptic components in the Merkel cell-neurite complex, a gentle-touch receptor. *Ann. NY Acad. Sci.* **1279,** 13–21 (2013).
6. Hartschuh, W. & Weihe, E. Fine structural analysis of the synaptic junction of Merkel cell-axon-complexes. *J. Invest. Dermatol.* **75,** 159–165 (1980).
7. Iggo, A. & Muir, A. R. The structure and function of a slowly adapting touch corpuscle in hairy skin. *J. Physiol. (Lond.)* **200,** 763–796 (1969).
8. Woodbury, C. J. & Koerber, H. R. Central and peripheral anatomy of slowly adapting type I low-threshold mechanoreceptors innervating trunk skin of neonatal mice. *J. Comp. Neurol.* **505,** 547–561 (2007).
9. Wellnitz, S. A., Lesniak, D. R., Gerling, G. J. & Lumpkin, E. A. The regularity of sustained firing reveals two populations of slowly adapting touch receptors in mouse hairy skin. *J. Neurophysiol.* **103,** 3378–3388 (2010).

10. Piskorowski, R., Haeberle, H., Panditrao, M. V. & Lumpkin, E. A. Voltage-activated ion channels and Ca^{2+}-induced Ca^{2+} release shape Ca^{2+} signaling in Merkel cells. *Pflugers Arch.* **457,** 197–209 (2008).
11. Chalfie, M. Neurosensory mechanotransduction. *Nature Rev. Mol. Cell Biol.* **10,** 44–52 (2009).
12. Hu, J. & Lewin, G. R. Mechanosensitive currents in the neurites of cultured mouse sensory neurones. *J. Physiol. (Lond.)* **577,** 815–828 (2006).
13. Coste, B. *et al.* Piezo1 and Piezo2 are essential components of distinct mechanically activated cation channels. *Science* **330,** 55–60 (2010).
14. Walker, R. G., Willingham, A. T. & Zuker, C. S. A *Drosophila* mechanosensory transduction channel. *Science* **287,** 2229–2234 (2000).
15. Woo, S. H. *et al.* Piezo2 is required for Merkel cell mechanotransduction. *Nature* http://dx.doi.org/10.1038/nature13251(this issue).
16. Lesniak, D. R. *et al.* Computation identifies structural features that govern neuronal firing properties in slowly adapting touch receptors. *eLife* **3,** e01488 (2014).
17. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neurosci.* **8,** 1263–1268 (2005).
18. Taniguchi, H. *et al.* A resource of Cre driver lines for genetic targeting of GABAergic neurons in cerebral cortex. *Neuron* **71,** 995–1013 (2011).
19. Madisen, L. *et al.* A toolbox of Cre-dependent optogenetic transgenic mice for light-induced activation and silencing. *Nature Neurosci.* **15,** 793–802 (2012).
20. Quiroga, R. Q., Nadasdy, Z. & Ben-Shaul, Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput.* **16,** 1661–1687 (2004).
21. Dassule, H. R., Lewis, P., Bei, M., Maas, R. & McMahon, A. P. Sonic hedgehog regulates growth and morphogenesis of the tooth. *Development* **127,** 4775–4785 (2000).
22. Han, X. *et al.* A high-light sensitivity optical neural silencer: development and application to optogenetic control of non-human primate cortex. *Front. Syst. Neurosci.* **5,** 18 (2011).
23. Morrison, K. M., Miesegaes, G. R., Lumpkin, E. A. & Maricich, S. M. Mammalian Merkel cells are descended from the epidermal lineage. *Dev. Biol.* **336,** 76–83 (2009).
24. Maricich, S. M. *et al.* Merkel cells are essential for light-touch responses. *Science* **324,** 1580–1582 (2009).
25. Meyers, J. R. *et al.* Lighting up the senses: FM1–43 loading of sensory cells through nonselective ion channels. *J. Neurosci.* **23,** 4054–4065 (2003).
26. Koltzenburg, M., Stucky, C. L. & Lewin, G. R. Receptive properties of mouse sensory neurons innervating hairy skin. *J. Neurophysiol.* **78,** 1841–1850 (1997).
27. Johnson, K. O. & Lamb, G. D. Neural mechanisms of spatial tactile discrimination: neural patterns evoked by braille-like dot patterns in the monkey. *J. Physiol. (Lond.)* **310,** 117–144 (1981).
28. Phillips, J. R. & Johnson, K. O. Tactile spatial resolution. II. Neural representation of Bars, edges, and gratings in monkey primary afferents. *J. Neurophysiol.* **46,** 1192–1203 (1981).
29. Maricich, S. M., Morrison, K. M., Mathes, E. L. & Brewer, B. M. Rodents rely on Merkel cells for texture discrimination tasks. *J. Neurosci.* **32,** 3296–3300 (2012).
30. Yamashita, Y. & Ogawa, H. Slowly adapting cutaneous mechanoreceptor afferent units associated with Merkel cells in frogs and effects of direct currents. *Somatosens. Mot. Res.* **8,** 87–95 (1991).

Acknowledgements Thanks to R. Axel, A. MacDermott and the Lumpkin laboratory for helpful discussions and to D. Florez and R. Piskorowski for advice on whole-cell recordings. Funding was provided by NIH/NIAMS grants R01AR051219 and R21AR062307 (to E.A.L.), R01DE022358 (to A.P.) and fellowships to S.M. (5T32HL087745-05 and NIH/NINDS F32NS080544), M.N. (JSPS Research Fellowships for Young Scientists 24-7585), and A.M.N. (McNair Foundation). Microscopy and flow cytometry was performed with core support from the Columbia SDRC (P30AR044535) and Cancer Center (P30CA013696). Initial studies were performed at Baylor College of Medicine with assistance from Flow Cytometry and Genetically Engineered Mouse Shared Resources (P30CA125123).

Author Contributions S.M. screened transgenic mouse lines, and performed and analysed all *ex vivo* optogenetic experiments (Fig. 2 and Extended Data Figs 3–6). M.N. performed and analysed all whole-cell recordings (Fig. 1, Extended Data Fig. 1a, b and Extended Data Table 1). Y.B. performed and analysed recordings from *Atoh1* and *Piezo2* strains (Fig. 3, Extended Data Fig. 8 and Extended Data Table 2). A.M.N. performed qRT-PCR (Fig. 1i) and calcium imaging (Extended Data Fig. 1c–i). K.L.M. performed immunohistochemistry in *Atoh1* strains (Extended Data Fig. 7) and assisted in preparation of all figures. S.A.W. and E.A.L. conceived optogenetic strategies. P.F. generated initial ChR2 transgenic mouse lines. E.A.L. conceived and supervised the project. During this manuscript's peer-review process, we entered into a collaboration with S.H.W., S.R. and A.P., to analyse unpublished $Piezo2^{CKO}$ mice. S.R. generated $Piezo2^{flox/flox}$ mice and S.-H.W. generated and validated $Krt14^{Cre};Piezo2^{flox/flox}$ mice in the laboratory of A.P. The manuscript was written by S.M., M.N., Y.B. and E.A.L. and edited by A.M.N., K.L.M., S.A.W., P.F., S.-H.W. and A.P.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.A.L. (eal2166@columbia.edu).

LETTER

doi:10.1038/nature13251

Piezo2 is required for Merkel-cell mechanotransduction

Seung-Hyun Woo[1], Sanjeev Ranade[1], Andy D. Weyer[2], Adrienne E. Dubin[1], Yoshichika Baba[3], Zhaozhu Qiu[1,4], Matt Petrus[4], Takashi Miyamoto[1]†, Kritika Reddy[4], Ellen A. Lumpkin[3], Cheryl L. Stucky[2] & Ardem Patapoutian[1]

How we sense touch remains fundamentally unknown[1,2]. The Merkel cell–neurite complex is a gentle touch receptor in the skin that mediates slowly adapting responses of Aβ sensory fibres to encode fine details of objects[3–6]. This mechanoreceptor complex was recognized to have an essential role in sensing gentle touch nearly 50 years ago[3,4]. However, whether Merkel cells or afferent fibres themselves sense mechanical force is still debated, and the molecular mechanism of mechanotransduction is unknown[1,2,7–12]. Synapse-like junctions are observed between Merkel cells and associated afferents[6,13–15], and yet it is unclear whether Merkel cells are inherently mechanosensitive or whether they can rapidly transmit such information to the neighbouring nerve[1,2,16,17]. Here we show that Merkel cells produce touch-sensitive currents *in vitro*. Piezo2, a mechanically activated cation channel, is expressed in Merkel cells. We engineered mice deficient in Piezo2 in the skin, but not in sensory neurons, and show that Merkel-cell mechanosensitivity completely depends on Piezo2. In these mice, slowly adapting responses *in vivo* mediated by the Merkel cell–neurite complex show reduced static firing rates, and moreover, the mice display moderately decreased behavioural responses to gentle touch. Our results indicate that Piezo2 is the Merkel-cell mechanotransduction channel and provide the first line of evidence that Piezo channels have a physiological role in mechanosensation in mammals. Furthermore, our data present evidence for a two-receptor-site model, in which both Merkel cells and innervating afferents act together as mechanosensors. The two-receptor system could provide this mechanoreceptor complex with a tuning mechanism to achieve highly sophisticated responses to a given mechanical stimulus[15,18,19].

We recently discovered Piezo proteins as an evolutionarily conserved mechanically activated (MA) cation channel family[20,21]. *Drosophila melanogaster* Piezo and zebrafish Piezo2b have been shown to be involved in somatosensory mechanotransduction[22,23]. Of the two mammalian Piezo members, Piezo1 and Piezo2, Piezo2 is expressed in dorsal root ganglion (DRG) sensory neurons and is required for a subset of mechanically activated currents in DRGs[20]. Here we focused on whether Piezo2 also has a role in somatosensory mechanotransduction in mammalian skin.

We generated a *Piezo2-GFP-IRES-Cre* knock-in reporter mouse line to detect Piezo2 expression *in vivo*. The *Piezo2-GFP-IRES-Cre* (*Piezo2*^{GFP}) allele contains enhanced green fluorescent protein (EGFP) fused to the carboxy (C)-terminal end of the *Piezo2* coding region, followed by Cre recombinase expressed through an internal ribosome entry site (IRES) (Fig. 1a). Mice carrying this allele express Piezo2–GFP fusion protein as well as Cre recombinase driven by the endogenous *Piezo2* promoter. Expression of the Piezo2–GFP fusion protein in human embryonic kidney (HEK293T) cells gives rise to mechanically activated currents indistinguishable from wild-type Piezo2-dependent currents (not shown). Using the *Piezo2–GFP* portion of the construct as a Piezo2 reporter, we examined Piezo2 expression in DRGs isolated from *Piezo2*^{GFP} mice as a positive control tissue[20]. When we co-stained using anti-GFP and anti-Piezo2 antibodies, GFP and Piezo2 expression patterns overlapped (Extended Data Fig. 1).

We examined both hairy and glabrous skin of *Piezo2*^{GFP} mice for Piezo2 expression. *Piezo2* was previously shown to be present at low levels in the skin by quantitative polymerase chain reaction with reverse transcription (qRT–PCR)[20], and here we found that GFP was specifically expressed in Merkel cells (~0.05–0.1% of total epithelial cells from dorsal skin) within whisker pad, dorsal skin, and foot pad (Fig. 1b–f and Extended Data Fig. 2a–c). We used antibodies against keratin 8 (Krt8, a marker for Merkel cells) and neurofilament heavy polypeptide (Nefh, a marker for myelinated sensory afferents) in conjunction with GFP antibody to visualize the precise localization of Piezo2 within Merkel cell–neurite complexes. GFP was expressed in Merkel cells, preferentially on the side adjacent to afferent fibre innervation (Fig. 1b–f and Extended Data Fig. 2d–h). GFP was also present in Nefh$^+$ sensory afferents, including the fibres that innervated Merkel cells (Fig. 1c, d and Extended Data Fig. 2d–h).

Because of the close proximity between Merkel cells and innervating afferents, it was difficult to convincingly conclude that GFP and Piezo2 were indeed present in Merkel cells. Therefore, we used atonal homologue 1 fused to GFP (*Atoh1*^{GFP}) reporter mice expressing GFP in Merkel cells to purify these cells and perform qRT–PCR analysis for *Piezo2* (ref. 24). GFP$^+$ and GFP$^-$ cells from *Atoh1*^{GFP} dorsal skin were purified by fluorescence-activated cell sorting (FACS), and total RNA from these samples was subjected to qRT–PCR for *Krt8*, *Piezo2* and keratin 14 (*Krt14*, a marker for basal keratinocytes and Merkel cells) (Fig. 1g, h). As expected, GFP$^+$ cells showed high expression levels of *Krt8* (Fig. 1h, left), indicating that they were indeed Merkel cells. Importantly, GFP$^+$ cells also showed high *Piezo2* levels comparable to DRG neurons, confirming *Piezo2* expression in Merkel cells (Fig. 1h, middle). *Piezo2* levels in GFP$^-$ cells were minimal, consistent with our GFP immunofluoresence results in *Piezo2*^{GFP} mouse skin as well as our previous qRT–PCR results from skin (Fig. 1b–f)[20]. As expected, *Krt14* expression was observed in both GFP$^+$ and GFP$^-$ epithelial cells, but not in DRGs. This is in agreement with GFP$^-$ epithelial cells being mainly comprised of basal keratinocytes (Fig. 1h, right).

To probe the role of Piezo2 in Merkel cells, we engineered skin-specific *Piezo2* knockout mice. We generated floxed *Piezo2* mice (*Piezo2*$^{fl/fl}$), which contained two *loxP* sites flanking exons 43 through 45 (Fig. 2a). We targeted this specific region close to the C terminus as it is highly conserved across different species, and moreover, the Cre excision of exons 43–45 causes a frameshift mutation in *Piezo2*, introducing an early stop codon (Extended Data Fig. 3). We crossed *Piezo2*$^{fl/fl}$ mice to *Krt14Cre* mice to generate *Krt14Cre;Piezo2*$^{fl/fl}$ conditional knockout (cKO) mice, in which Piezo2 expression is ablated in all epithelial cells including Merkel cells (Fig. 2a)[25].

Krt14Cre;Piezo2$^{fl/fl}$ conditional knockout mice developed normally, and their skin, including Merkel cell–neurite complexes, appeared normal compared to *Piezo2*$^{fl/fl}$ wild-type littermates (Extended Data Fig. 4). To study a *Piezo2* deletion specifically in Merkel cells, we further generated *Krt14Cre;Piezo2*$^{fl/fl}$;*Atoh1*^{GFP} mice. This particular mouse line allowed for the purification of GFP$^+$ Piezo2-deficient and wild-type Merkel cells for qRT–PCR (Fig. 2b, c). In GFP$^+$ Piezo2-deficient Merkel

[1]Howard Hughes Medical Institute, Molecular and Cellular Neuroscience, The Scripps Research Institute, La Jolla, California 92037, USA. [2]Departments of Cell Biology, Neurobiology, and Anatomy, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA. [3]Departments of Dermatology & Physiology and Cellular Biophysics, Columbia University, New York, New York 10032, USA. [4]Genomics Institute of the Novartis Research Foundation, San Diego, California 92121, USA. †Present address: Gladstone Institute of Neurological Disease, San Francisco, California 94158, USA.

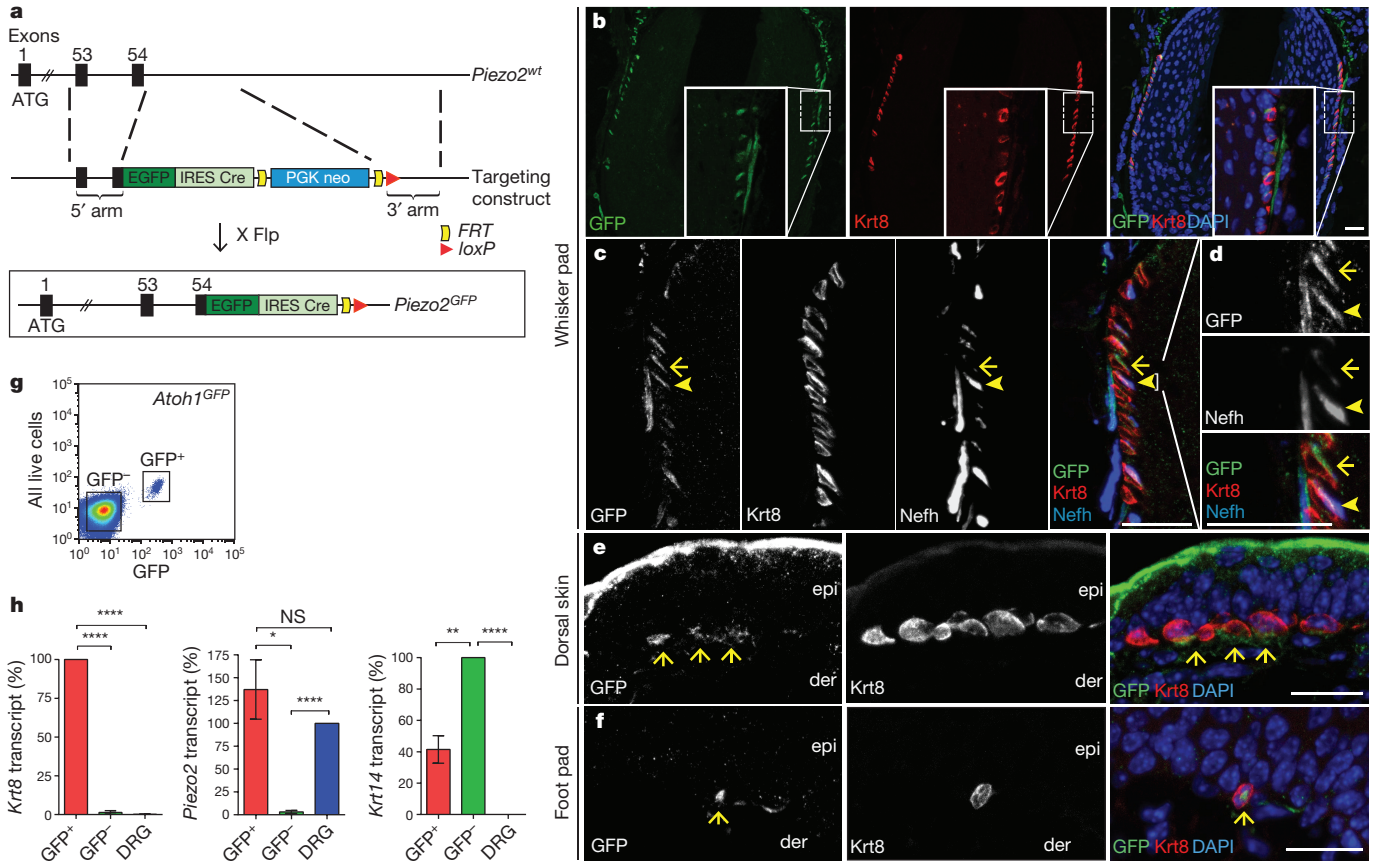


Figure 1 | Piezo2 expression in hairy and glabrous skin. **a**, A schematic diagram of the *Piezo2*^{GFP} allele generation. Flp, flippase. **b**, GFP and Krt8 co-staining in the whisker follicle at a lower magnification. **c**, **d**, GFP, Krt8, and Nefh co-staining in the whisker follicle at a higher magnification. Panel **d** shows a magnified view of the bracketed area in **c**. Arrowheads mark the co-localization of GFP, Krt8, and Nefh. Note that in areas where Nefh^{+} fibres are missing, GFP and Krt8 still co-localize (arrows). **e**, **f**, GFP and Krt8 co-staining in a touch dome (**e**) and in glabrous skin (**f**). Arrows mark the position of Krt8^{+} Merkel cells. Scale bars **b**–**f**, 20 μm. epi, epidermis; der, dermis. **g**, A representative fluorescence-activated cell sorting (FACS) plot (out of 12 experiments) of live epithelial cells isolated from *Atoh1*^{GFP} skin. **h**, qRT–PCR analysis ($n = 4$) of GFP^{+} and GFP^{-} cells and DRG neurons isolated from *Atoh1*^{GFP} mice. Bars represent mean ± s.e.m. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{****}P < 0.0001$; NS, not significantly different, unpaired *t*-test with Welch's correction.

cells, nearly 95% of *Piezo2* transcript was degraded compared to wild type, whereas *Piezo2* levels in DRGs from these knockout mice were normal (Fig. 2b, c). Moreover, immunofluorescence using the Piezo2 antibody in Piezo2-deficient whisker follicles revealed that Piezo2 expression was abolished in Merkel cells (Fig. 2d, e) but still intact in afferent fibres that innervated Merkel cells (Fig. 2e, f). These results indicate that our skin-specific *Piezo2* conditional knockout mice efficiently ablated Piezo2 only in the epithelial lineage, including Merkel cells.

Next, we asked whether Merkel cells are mechanosensitive, and if they are, whether their mechanosensitivity is dependent on Piezo2. Using *Krt14Cre;Piezo2*$^{fl/fl}$*;Atoh1*^{GFP} and littermate control mice, GFP^{+} wild-type and Piezo2-deficient Merkel cells were FACS-purified for whole-cell patch clamp electrophysiological recordings. When Merkel cells were stimulated by gentle poking using a blunt glass probe, wild-type cells ($n = 15$) responded robustly in a stimulus-dependent manner, producing rapidly adapting currents ($\tau = 6.4 \pm 1.5$ ms, Fig. 3a, left). However, none of the Piezo2-deficient Merkel cells ($n = 13$) showed mechanically activated currents (Fig. 3a, right). The lack of mechanosensitivity in Piezo2-deficient cells could not be accounted for by compromised membrane properties (Extended Data Table 1).

Mechanosensitive currents were rapidly adapting in voltage clamp mode. In current clamp mode, sustained depolarization was induced by gentle mechanical stimuli in all wild-type but not in Piezo2-deficient Merkel cells (Fig. 3b, c, top). Subsequent voltage clamp recordings in these same cells confirmed activation of rapidly adapting currents in wild-type cells upon mechanical stimulation (Fig. 3b, c, middle). Importantly, both wild-type and Piezo2-deficient Merkel cells were depolarized similarly by small current injections (Fig. 3b, c, bottom, and Extended Data Table 1). We next asked whether mechanically activated currents through Piezo2 could be responsible for the prolonged depolarization in wild-type cells (Fig. 3b, top). To answer this, we injected short-duration currents with amplitudes similar to Piezo2-dependent currents, and indeed observed sustained depolarization in wild-type Merkel cells (Extended Data Fig. 5), indicating that a predominant rapidly adapting current could account for the observed sustained receptor potentials in these cells[26]. Other conductances may also contribute to this prolonged depolarization, including voltage-gated calcium currents, which are reported to have a threshold near −20 mV (ref. 17), and stochastic gating of Merkel-cell-membrane ion channels, which could cause significant voltage fluctuations due to the high R_m of these cells (Extended Data Table 1). Thus, activation of Merkel cells by gentle touch produces a long-lasting depolarization that may ultimately contribute to the slowly adapting firing of SAI nerve fibres that innervate them. Collectively, these data show for the first time that Merkel cells are indeed mechanosensitive, and that Piezo2 is required in Merkel cells to produce their mechanical currents *in vitro*.

Next, we asked whether Piezo2 ablation in Merkel cells has any effect on slowly adapting responses of Aβ touch dome afferents. To answer this, we performed *ex vivo* skin-saphenous nerve recordings in *Krt14Cre;Piezo2*$^{fl/fl}$ (cKO) and *Piezo2*$^{fl/fl}$ (WT) littermates (Fig. 4). The overall proportions of Aβ fibre subtypes found by an electrical search stimulus were similar in both groups (Fig. 4a). Interestingly, firing

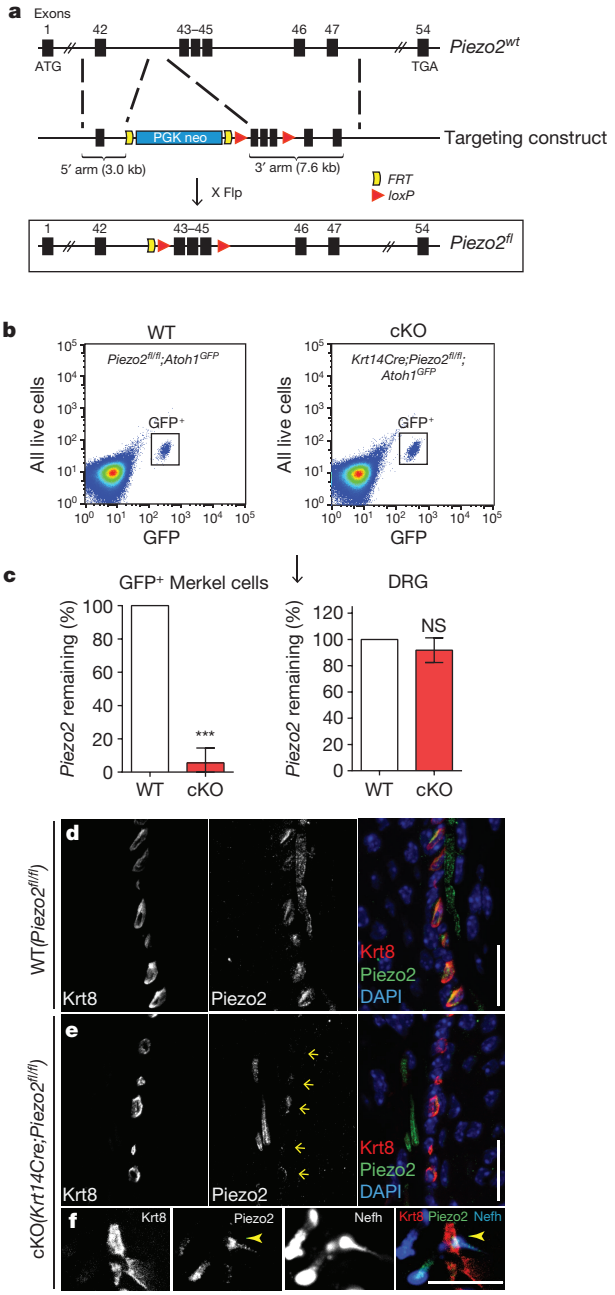


Figure 2 | Generation and characterization of skin-specific *Piezo2* conditional knockout mice. **a**, A schematic diagram of the *Piezo2*^{fl} allele generation. Flp, flippase. **b**, Representative FACS plots (out of 9 experiments) of live epithelial cells isolated from wild-type (WT) and conditional knockout (cKO) dorsal skin. **c**, qRT–PCR analysis ($n = 3$) showing *Piezo2* levels in GFP$^+$ cells and DRG neurons from wild-type (WT) and conditional knockout (cKO) mice. Bars represent mean ± s.e.m. ***$P < 0.001$; NS, not significantly different, unpaired *t*-test with Welch's correction. **d**, **e**, Piezo2 and Krt8 co-staining in wild-type (**d**) and conditional knockout (**e**) whisker pads. In **e**, arrows mark the position of Krt8$^+$ Merkel cells. **f**, Piezo2, Krt8 and Nefh co-staining in *Piezo2* conditional knockout whisker pad. Arrowheads mark the co-localization of Piezo2 and Nefh in sensory afferents. Scale bars **d**–**f**, 20 μm.

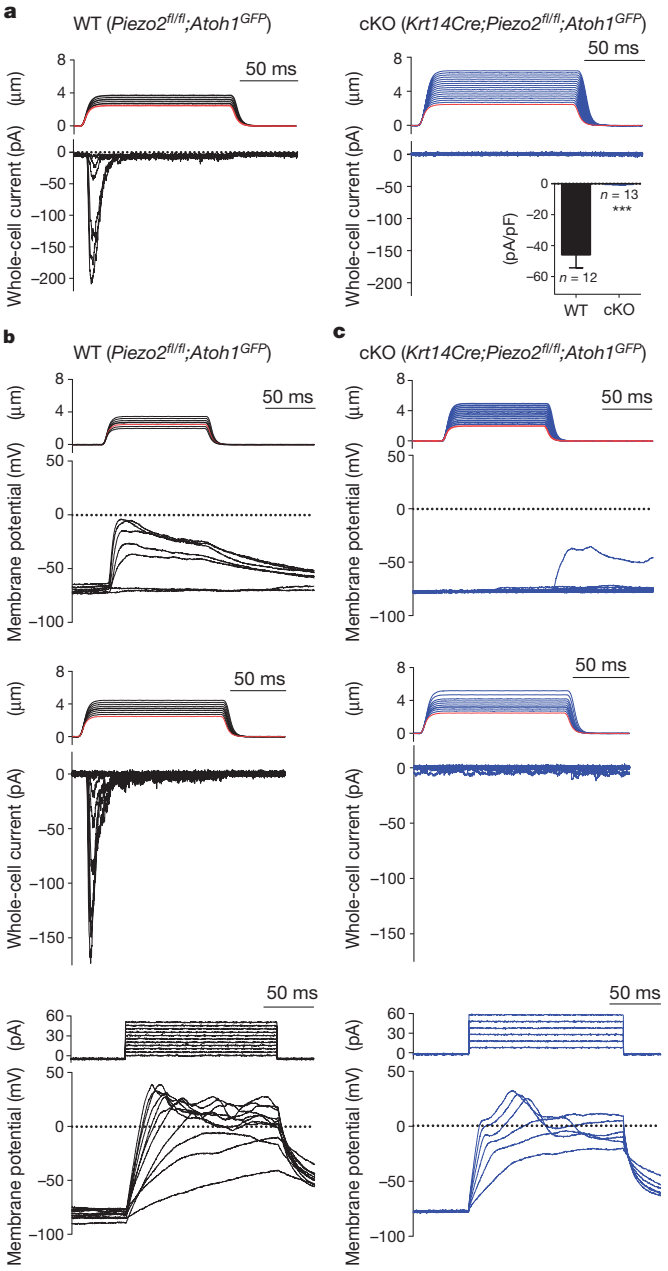


Figure 3 | Mechanically activated currents in Merkel cells depend on Piezo2. **a**, Representative traces of mechanically activated inward currents by a gentle poking stimulus in Merkel cells from wild-type (left) and conditional knockout (right) murine skin. The inset shows average mechanically activated whole-cell current density measured in wild-type and Piezo2-deficient Merkel cells (using CsCl intracellular solution). Error bars represent mean ± s.e.m. ***$P < 0.001$, Student's *t*-test. **b**, **c**, Representative traces (out of 6 experiments) of wild-type (**b**) and Piezo2-deficient (**c**) Merkel cells. Current clamp recordings displaying a change in membrane potential in response to gentle mechanical stimuli (top); subsequent displacement assays in voltage clamp mode (middle); current clamp recordings showing similar changes in membrane potential elicited by electrical stimulation (bottom) (using K-gluconate intracellular solution). Red lines in stimulus traces indicate the displacement at which the probe visibly touched the cell.

frequencies of overall slowly adapting Aβ fibres at various forces were reduced in *Piezo2* conditional knockout animals compared to wild-type littermates, but they were not abolished (Fig. 4b). Conduction velocity of these slowly adapting fibres was also slightly reduced in conditional knockout mice, compared to wild type, but von Frey thresholds were similar in both groups (Extended Data Fig. 6). In contrast, firing frequencies of RA Aβ fibres were comparable between Piezo2-deficient and wild-type mice (Extended Data Fig. 7 and Extended Data Table 2).

To specifically probe the effect of Piezo2 ablation on Merkel-cell-innervating SAI afferents, we used a targeted approach to record directly from touch dome afferents, which were identified by FM1-43 dye uptake[27]. Upon mechanical stimulation, touch dome afferents of *Piezo2* conditional knockout mice showed noticeably short, reduced firing, compared

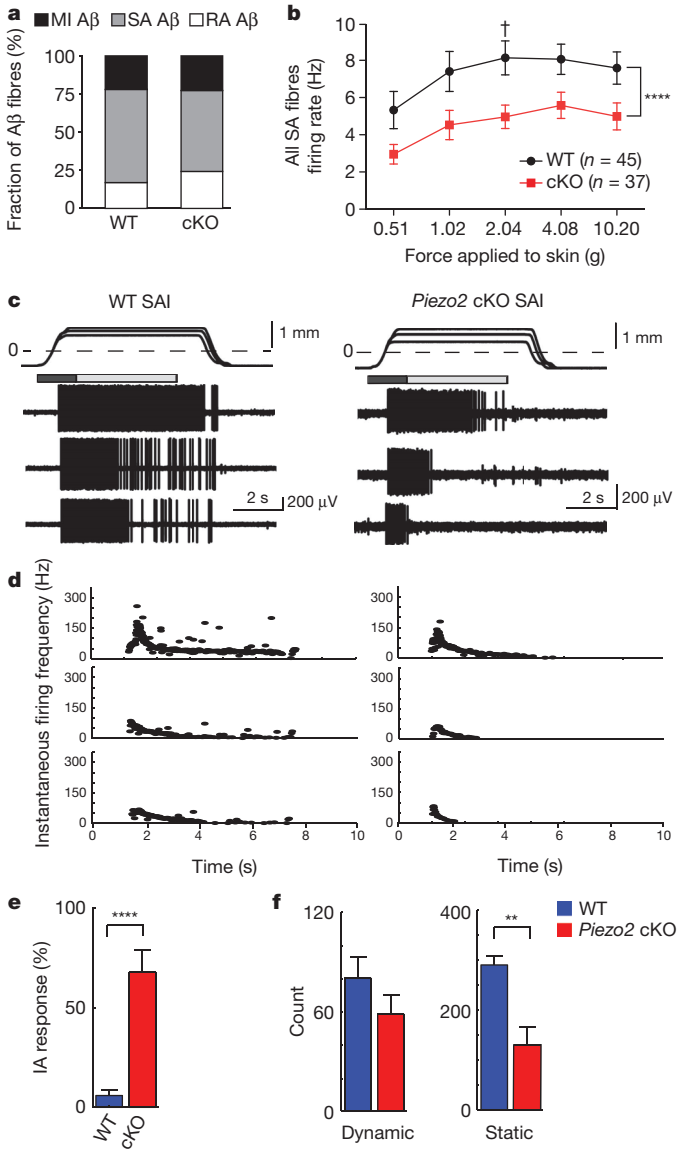


Figure 4 | *Ex vivo* skin-saphenous nerve recordings in wild-type and *Piezo2* conditional knockout mice. **a**, Proportions of mechanically insensitive (MI), slowly adapting (SA), and rapidly adapting (RA) Aβ fibres in wild-type ($n = 59$) and *Piezo2* conditional knockout mice ($n = 66$). The relatively low rapidly adapting Aβ fibre percentage in both groups in comparison to previous studies is probably owing to the genetic background of our animals[30]. **b**, Firing rates of all SA Aβ fibres at increasing forces in wild-type and *Piezo2* conditional knockout mice. $^{****}P < 0.0001$; $^{\dagger}P < 0.05$, two-way analysis of variance (ANOVA) with Bonferroni post-hoc analysis. **c**, Representative recordings from touch dome afferents in wild-type (left) vs *Piezo2* conditional knockout (right) mice. Top traces show ramp-and-hold displacements at three magnitudes with corresponding spike trains below. Dashed line marks the point of skin contact (0 mm). Boxes indicate dynamic (dark grey, 1.5 s after displacement command onset) and static phases for analysis (light grey, 4 s after the beginning of hold command). **d**, Instantaneous firing frequencies of the responses in (**c**). **e**, Proportion of intermediately adapting (IA) responses to supra-threshold displacements in touch dome afferents. $^{****}P < 0.0001$, Fisher's exact test. **f**, Maximum number of spikes in the dynamic and static response phases. $^{**}P \leq 0.01$, Student's *t*-test. Error bars represent mean ± s.e.m. For afferents in **c**–**f**, wild-type $n = 5$ and *Piezo2* conditional knockout $n = 6$. Note that for **a**, **b**, recordings were made using a force-controlled mechanical stimulator; for **c**–**f**, a displacement control was used for the directed recordings of FM1-43-labelled afferents.

to wild-type littermates (Fig. 4c, d and Extended Data Table 3). Unlike wild-type afferents that maintained firing for the duration of ramp-and-hold stimuli, touch dome afferents of Piezo2-deficient mice produced intermediately adapting firing patterns (Fig. 4e). The number of spikes fired (Fig. 4f and Extended Data Table 3) and overall mean firing rates were significantly diminished in touch dome afferents of Piezo2-deficient mice (17 ± 3 Hz; $n = 33$ displacements from 6 afferents), compared to wild type (27 ± 4 Hz; $n = 28$ displacements from 5 afferents). This difference was more pronounced during the static stimulation phase, resulting in reduced static firing rates (Fig. 4f). Together, our data suggest that Piezo2 is required for proper mechanosensory encoding in Merkel cell–neurite complexes in the intact skin.

We next performed somatosensory behavioural assays in *Piezo2* conditional knockout and control littermate mice. Because Merkel cell–neurite complexes are known to mediate gentle touch sensation, we subjected these animals to tests for gentle touch, texture and pain, which included von Frey filaments, cotton swab sensitivity of the hindpaw, texture discrimination assay and a Randall Selitto test. For most of the above-mentioned assays, *Piezo2* conditional knockout animals behaved similarly to control littermates (data not shown). In the automated von Frey filament test, Piezo2-deficient mice showed a significant decrease in per cent paw withdrawal response only at lower forces (1.0 g–1.5 g), but showed comparable responses to controls at higher forces (Fig. 5a). At low forces, not only were the responses less frequent, but fewer knockout mice responded (Fig. 5b). These results indicate that Piezo2-deficient mice are affected in sensing only low force mechanical stimuli. This phenotype is in agreement with our skin–nerve recordings, which show a reduction, but not complete ablation of slowly adapting responses in Piezo2-deficient mice.

We have provided answers to critical questions regarding the functionality of the Merkel cell–neurite complex. Our study is the first, to our knowledge, to show that Merkel cells display touch-sensitive currents, and that Piezo2 is required for Merkel-cell mechanotransduction. This represents the first definitive evidence for a mammalian Piezo family member to be involved in mechanotransduction *in vivo*. Moreover, we show that Merkel cells have a partial role in the generation of slowly adapting responses and in gentle touch perception in mice. Interestingly, the accompanying manuscript[28] demonstrates that *Krt14Cre;Atoh1*^{CKO} animals, which completely lack epidermal Merkel cells in their skin, show similar SAI firing deficits as *Piezo2* conditional knockout mice in skin–nerve recordings. This reinforces the hypothesis that Piezo2 is the principal Merkel-cell mechanotransduction channel *in vivo*. Our current data are most consistent with a two-receptor-site model, which proposes that both Merkel cells and Aβ sensory afferents are necessary for mediating proper mechanically activated slowly adapting responses[15,18,19,29]. This model explains why slowly adapting fibres are only partially affected when Merkel cells are present but not mechanosensitive. The observed expression of Piezo2 in the afferents that innervate Merkel cells may also support this model. Analysis of the sensory afferent-specific Piezo2-deficient animal model will provide further clues to reveal the extent to which mechanotransduction is directly dependent on the nerves. Most peripheral sensory

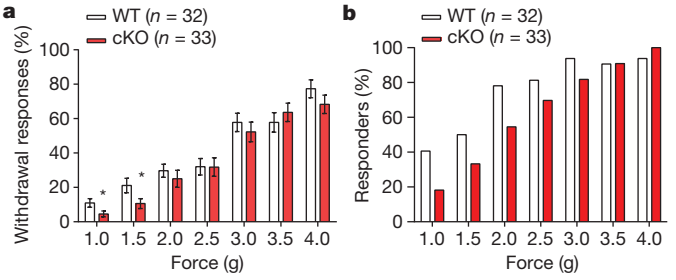


Figure 5 | Gentle touch assay in wild-type and *Piezo2* conditional knockout mice. **a**, Per cent paw withdrawal responses to gentle von Frey filament stimulation between 1.0 and 4.0 g forces in wild-type and *Piezo2* conditional knockout mice. Error bars represent the mean ± s.e.m. $^{*}P < 0.05$, Student's *t*-test. **b**, Per cent of mice responding to von Frey filament stimulation in wild-type and *Piezo2* conditional knockout mice.

receptors are thought to have evolved with a single receptor site for sensory transduction. The Merkel cell–neurite complex is known to have the highest spatial resolution among other cutaneous mechanoreceptors, and this allows for deciphering fine spatial details such as shape, edge and curvature[5]. We speculate that mechanosensors present both at nerve endings and in Merkel cells act together to convey the exquisite mechanosensitivity of this complex[15,18,19].

METHODS SUMMARY

Detailed information regarding mouse lines, immunofluorescence, Merkel-cell isolation, qRT–PCR, FACS, whole-cell electrophysiology, skin-nerve recordings and behavioural assays is provided in the Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 November 2013; accepted 14 March 2014.
Published online 6 April 2014.

1. Abraira, V. E. & Ginty, D. D. The sensory neurons of touch. *Neuron* **79,** 618–639 (2013).
2. Maksimovic, S., Baba, Y. & Lumpkin, E. A. Neurotransmitters and synaptic components in the Merkel cell-neurite complex, a gentle-touch receptor. *Ann. NY Acad. Sci.* **1279,** 13–21 (2013).
3. Chambers, M. R. & Iggo, A. Slowly-adapting cutaneous mechanoreceptors. *J. Physiol. (Lond.)* **192,** (suppl.) 26P–27P (1967).
4. Iggo, A. & Muir, A. R. The structure and function of a slowly adapting touch corpuscle in hairy skin. *J. Physiol. (Lond.)* **200,** 763–796 (1969).
5. Johnson, K. O. The roles and functions of cutaneous mechanoreceptors. *Curr. Opin. Neurobiol.* **11,** 455–461 (2001).
6. Halata, Z., Grim, M. & Bauman, K. I. Friedrich Sigmund Merkel and his "Merkel cell", morphology, development, and physiology: review and new results. *Anat. Rec. A Discov. Mol. Cell. Evol. Biol.* **271A,** 225–239 (2003).
7. Ikeda, I., Yamashita, Y., Ono, T. & Ogawa, H. Selective phototoxic destruction of rat Merkel cells abolishes responses of slowly adapting type I mechanoreceptor units. *J. Physiol. (Lond.)* **479,** 247–256 (1994).
8. Mills, L. R. & Diamond, J. Merkel cells are not the mechanosensory transducers in the touch dome of the rat. *J. Neurocytol.* **24,** 117–134 (1995).
9. Senok, S. S., Baumann, K. I. & Halata, Z. Selective phototoxic destruction of quinacrine-loaded Merkel cells is neither selective nor complete. *Exp. Brain Research* **110,** 325–334 (1996).
10. Kinkelin, I., Stucky, C. L. & Koltzenburg, M. Postnatal loss of Merkel cells, but not of slowly adapting mechanoreceptors in mice lacking the neurotrophin receptor p75. *Eur. J. Neurosci.* **11,** 3963–3969 (1999).
11. Maricich, S. M. *et al.* Merkel cells are essential for light-touch responses. *Science* **324,** 1580–1582 (2009).
12. Maricich, S. M., Morrison, K. M., Mathes, E. L. & Brewer, B. M. Rodents rely on Merkel cells for texture discrimination tasks. *J. Neuroscience* **32,** 3296–3300 (2012).
13. Hartschuh, W. & Weihe, E. Fine structural analysis of the synaptic junction of Merkel cell-axon-complexes. *J. Invest. Dermatol.* **75,** 159–165 (1980).
14. Gu, J., Polak, J. M., Tapia, F. J., Marangos, P. J. & Pearse, A. G. Neuron-specific enolase in the Merkel cells of mammalian skin. The use of specific antibody as a simple and reliable histologic marker. *Am. J. Pathol.* **104,** 63–68 (1981).
15. Fagan, B. M. & Cahusac, P. M. Evidence for glutamate receptor mediated transmission at mechanoreceptors in the skin. *Neuroreport* **12,** 341–347 (2001).
16. Diamond, J., Holmes, M. & Nurse, C. A. Are Merkel cell-neurite reciprocal synapses involved in the initiation of tactile responses in salamander skin? *J. Physiol. (Lond.)* **376,** 101–120 (1986).
17. Yamashita, Y., Akaike, N., Wakamori, M., Ikeda, I. & Ogawa, H. Voltage-dependent currents in isolated single Merkel cells of rats. *J. Physiol. (Lond.)* **450,** 143–162 (1992).
18. Cahusac, P. M. & Mavulati, S. C. Non-competitive metabotropic glutamate 1 receptor antagonists block activity of slowly adapting type I mechanoreceptor units in the rat sinus hair follicle. *Neuroscience* **163,** 933–941 (2009).
19. Press, D., Mutlu, S. & Guclu, B. Evidence of fast serotonin transmission in frog slowly adapting type 1 responses. *Somatosens. Mot. Res.* **27,** 174–185 (2010).
20. Coste, B. *et al.* Piezo1 and Piezo2 are essential components of distinct mechanically activated cation channels. *Science* **330,** 55–60 (2010).
21. Coste, B. *et al.* Piezo proteins are pore-forming subunits of mechanically activated channels. *Nature* **483,** 176–181 (2012).
22. Kim, S. E., Coste, B., Chadha, A., Cook, B. & Patapoutian, A. The role of *Drosophila* Piezo in mechanical nociception. *Nature* **483,** 209–212 (2012).
23. Faucherre, A., Nargeot, J., Mangoni, M. E. & Jopling, C. *piezo2b* regulates vertebrate light touch response. *J. Neurosci.* **33,** 17089–17094 (2013).
24. Rose, M. F. *et al. Math1* is essential for the development of hindbrain neurons critical for perinatal breathing. *Neuron* **64,** 341–354 (2009).
25. Dassule, H. R., Lewis, P., Bei, M., Maas, R. & McMahon, A. P. Sonic hedgehog regulates growth and morphogenesis of the tooth. *Development* **127,** 4775–4785 (2000).
26. Lesniak, D. R. *et al.* Computation identifies structural features that govern neuronal firing properties in slowly adapting touch receptors. *eLife* **3,** e01488 (2014).
27. Wellnitz, S. A., Lesniak, D. R., Gerling, G. J. & Lumpkin, E. A. The regularity of sustained firing reveals two populations of slowly adapting touch receptors in mouse hairy skin. *J. Neurophysiol.* **103,** 3378–3388 (2010).
28. Maksimovic, S. *et al.* Epidermal Merkel cells are mechanosensory cells that tune mammalian touch receptors. *Nature* http://dx.doi.org/10.1038/nature13250 (this issue).
29. Yamashita, Y. & Ogawa, H. Slowly adapting cutaneous mechanoreceptor afferent units associated with Merkel cells in frogs and effects of direct currents. *Somatosens. Mot. Res.* **8,** 87–95 (1991).
30. Milenkovic, N., Wetzel, C., Moshourab, R. & Lewin, G. R. Speed and temperature dependences of mechanotransduction in afferent fibers recorded from the mouse saphenous nerve. *J. Neurophysiol.* **100,** 2771–2783 (2008).

Acknowledgements We would like to thank S. Murthy and B. Coste for suggestions. Research was supported by the Howard Hughes Medical Institute (to A.P.) and NIH grants R01DE022358 (to A.P.) and R01AR051219 (to E.A.L.).

Author Contributions S.-H.W. conducted experiments for Figs 1–3 and Extended Data Figs 1–5. A.E.D. conducted and analysed *in vitro* electrophysiology experiments in Fig. 3 and Extended Data Fig. 5. A.D.W. performed and analysed general skin–nerve recordings in Fig. 4a, b and Extended Data Figs 6 and 7. M.P. and K.R. performed behavioural experiments in Fig. 5. S.R. generated *Piezo2*$^{fl/fl}$ mice. Z.Q. and T.M. generated *Piezo2*^{GFP} mice. During this manuscript's peer-review process, we entered into a collaboration with Y.B. and E.A.L. E.A.L. conceived, and Y.B. performed and analysed targeted skin–nerve recordings in Fig. 4c–f. E.A.L. and C.L.S. contributed to the editing of the manuscript. S.-H.W. and A.P. designed experiments and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.P. (ardem@scripps.edu).

LETTER

doi:10.1038/nature13169

Scalable control of mounting and attack by Esr1$^+$ neurons in the ventromedial hypothalamus

Hyosang Lee[1,2], Dong-Wook Kim[3], Ryan Remedios[1], Todd E. Anthony[1], Angela Chang[1], Linda Madisen[4], Hongkui Zeng[4] & David J. Anderson[1,2,3]

Social behaviours, such as aggression or mating, proceed through a series of appetitive and consummatory phases[1] that are associated with increasing levels of arousal[2]. How such escalation is encoded in the brain, and linked to behavioural action selection, remains an unsolved problem in neuroscience. The ventrolateral subdivision of the murine ventromedial hypothalamus (VMHvl) contains neurons whose activity increases during male–male and male–female social encounters. Non-cell-type-specific optogenetic activation of this region elicited attack behaviour, but not mounting[3]. We have identified a subset of VMHvl neurons marked by the oestrogen receptor 1 (Esr1), and investigated their role in male social behaviour. Optogenetic manipulations indicated that Esr1$^+$ (but not Esr1$^-$) neurons are sufficient to initiate attack, and that their activity is continuously required during ongoing agonistic behaviour. Surprisingly, weaker optogenetic activation of these neurons promoted mounting behaviour, rather than attack, towards both males and females, as well as sniffing and close investigation. Increasing photostimulation intensity could promote a transition from close investigation and mounting to attack, within a single social encounter. Importantly, time-resolved optogenetic inhibition experiments revealed requirements for Esr1$^+$ neurons in both the appetitive (investigative) and the consummatory phases of social interactions. Combined optogenetic activation and calcium imaging experiments *in vitro*, as well as c-Fos analysis *in vivo*, indicated that increasing photostimulation intensity increases both the number of active neurons and the average level of activity per neuron. These data suggest that Esr1$^+$ neurons in VMHvl control the progression of a social encounter from its appetitive through its consummatory phases, in a scalable manner that reflects the number or type of active neurons in the population.

To identify molecular markers for neurons that mediate aggression, we performed double-labelling experiments using markers for subsets of neurons in VMHvl[4], and the neuronal activation marker c-Fos[5], in resident males that had recently attacked an intruder. These studies identified Esr1 (ref. 6) as enriched in cells activated during aggression (>80% of c-Fos$^+$ cells Esr1$^+$; Extended Data Fig. 1h–v). To gain genetic access to these neurons, we generated a knock-in mouse line in which the *Cre* recombinase gene was targeted to the 3′ end of the *Esr1* coding sequence in a gene-conserving manner (Fig. 1a, b). *In situ* hybridization for *Cre* messenger RNA revealed an expression pattern similar to that of *Esr1* mRNA (Fig. 1c–h). As in wild-type mice[7], the expression of *Esr1-Cre* mRNA in VMHvl was higher in females than in males (Fig. 1g–j and Extended Data Fig. 1a–d). Anti-Esr1 antibody staining (Fig. 1i, j, s, u) indicated that the fraction of Esr1$^+$ cells (~40%; see below) was similar in wild-type and gene-targeted mice.

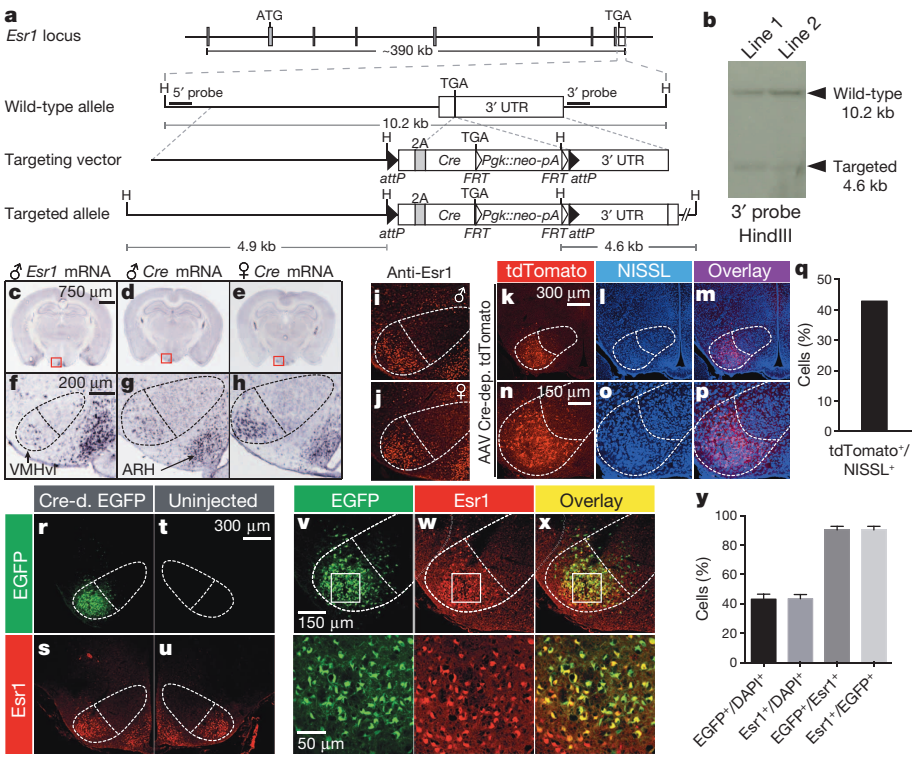


Figure 1 | Generation and characterization of a knock-in mouse line expressing Cre recombinase in Esr1$^+$ cells. **a**, Strategy for targeting the *Esr1* locus. H, HindIII; 3′ UTR, 3′ untranslated region; 2A, F2A sequence; Pgk, phosphoglycerate kinase promoter; *neo*, neomycin-resistance gene; pA, polyadenylation signal. **b**, Southern blot of HindIII-digested genomic DNA from two correctly targeted *Esr1*$^{cre/+}$ embryonic stem cell lines. Wild-type (10.2 kb) and targeted (4.6 kb) alleles are revealed by a 3′ probe (**a**). **c**–**h**, *In situ* hybridization for *Esr1* mRNA in wild-type male (**c**, **f**, ©2014 Allen Institute for Brain Science. Allen Mouse Brain Atlas [Internet]. Available from: http://mouse.brain-map.org/experiment/show/79591677, Bregma −1.75 mm) and for Cre mRNA in *Esr1*$^{cre/+}$ male (**d**, **g**) and female (**e**, **h**) mice (Bregma −1.65 mm). VMHvl, ventrolateral subdivision of the ventromedial hypothalamus; ARH, arcuate nucleus. Dotted outline indicates VMH. **i**–**x**, Immunostaining for Esr1 protein (red) in wild-type (**i**, male; **j**, female) and *Esr1*$^{cre/+}$ female mice (**s**, **u**, **w**). **k**–**x**, Native fluorescence of Cre-dependent AAV-encoded markers in *Esr1*$^{cre/+}$ male (**k**–**p**, tdTomato) and female (**r**–**x**, EGFP) mice. Bottom panels in **v**–**x** are the boxed areas in top panels. **q**, **y**. Quantification of **k**–**p** (**q**, $n = 1$) and **r**–**x** (**y**, $n = 4$). Data are mean ± s.e.m. n = number of animals in this and all figures unless otherwise indicated.

[1]Division of Biology and Biological Engineering 156-29, California Institute of Technology, Pasadena, California 91125, USA. [2]Howard Hughes Medical Institute, Pasadena, California 91125, USA. [3]Computation and Neural Systems, California Institute of Technology, Pasadena, California 91125, USA. [4]Allen Institute for Brain Science, Seattle, Washington 98103, USA.

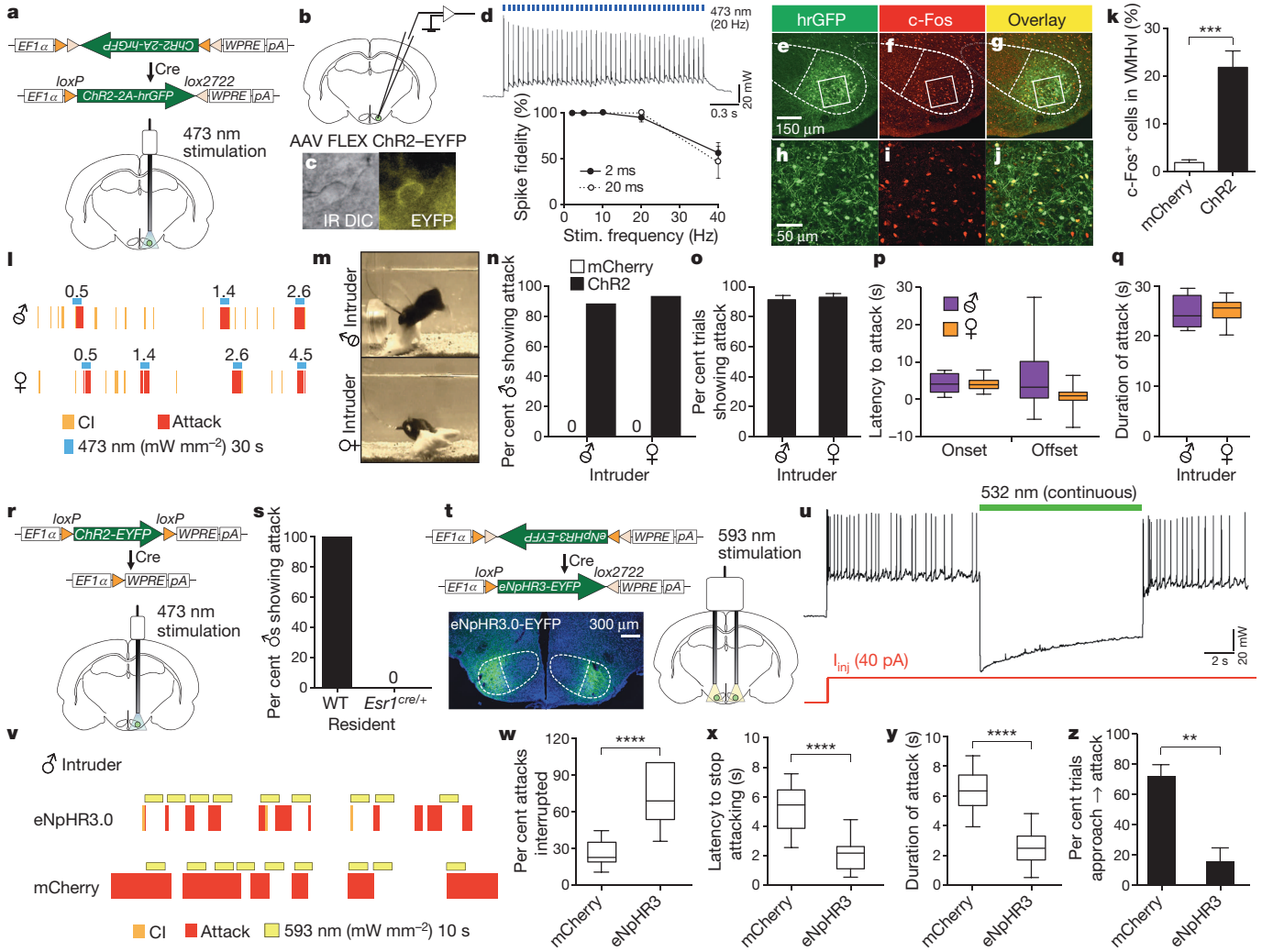


Figure 2 | Esr1$^+$ cells in VMHvl are necessary and sufficient for aggression. **a**, Strategy for optogenetic activation of Esr1$^+$ cells in VMHvl. EF1α, elongation factor 1α promoter; ChR2 is V5 epitope-tagged. **b**–**d**, Whole-cell patch-clamp recording from Esr1$^+$ cells in VMHvl (**c**, EYFP$^+$ cell) in acute hypothalamic slices. Photostimulation-evoked spiking (**d**, top) and quantification of spike fidelity (**d**, bottom) are shown (filled circles, 2 ms light pulse-width, $n = 11$ cells; open circles, 20 ms pulse-width, $n = 5$ cells). **e**–**j**, Double-labelling for Cre-dependent hrGFP viral reporter (**e**, **h**) and c-Fos (**f**, **i**) in VMHvl of *Esr1*$^{cre/+}$ males following photostimulation; **h**–**j**, boxed areas indicated in **e**–**g**. **k**, Quantification of **e**–**j** (mCherry, $n = 5$; ChR2, $n = 10$; ***$P < 0.001$; Mann–Whitney U-test). **l**, **m**, Representative raster plots (**l**) and video stills (**m**) illustrating photostimulation-evoked (blue bars; mW mm^{-2}) attack (**l**, red) or close investigation (CI, yellow) by ChR2-expressing *Esr1*$^{cre/+}$ males (**m**, black mice), towards a castrated male (strikethrough male symbol; **l**, **m**, upper) or an intact female (**l**, **m**, lower). See Supplementary Video 1. **n**–**q**, Quantification of attack parameters towards castrated males (ChR2, **n**, $n = 33$; **o**, $n = 23$; **p**, **q**, $n = 11$; mCherry, **n**, $n = 14$) or females (ChR2, **n**, $n = 28$; **o**, $n = 22$; **p**, **q**, $n = 16$; mCherry, **n**, $n = 10$). **r**, Photoactivation of Esr1$^-$ cells using 'Cre-out' ChR2 AAV. **s**, Percentage of wild-type (WT; $n = 4$) and *Esr1*$^{cre/+}$ ($n = 9$) males showing photostimulated attack towards castrated males (see Extended Data Fig. 3b, c). **t**, Expression of eNpHR3.0 in VMHvl Esr1$^+$ neurons. **u**, Whole-cell patch-clamp recording in acute hypothalamic slices, showing photostimulation-induced suppression of current injection-evoked spiking in eNpHR3.0-mCherry expressing Esr1$^+$ cells. **v**, Representative raster plots illustrating effect of photostimulation on attack towards a male intruder. **w**–**y**, Quantification of behavioural parameters (mCherry, $n = 10$; eNpHR3.0, $n = 13$; ****$P < 0.0001$; Mann–Whitney U-test). **z**, Percentage of photostimulation trials in which approach to intruder led to attack (mCherry, $n = 7$; eNpHR3, $n = 4$; **$P < 0.01$; Mann–Whitney U-test). Data are mean ± s.e.m. (**d**, **k**, **o**, **z**) or median ± min and maximum values (**p**, **q**, **w**–**y**).

Stereotaxic injection of recombinant adeno-associated viruses (AAVs) encoding Cre-dependent reporters into VMHvl of *Esr1*$^{cre/+}$ mice yielded marker-positive cells at a frequency (43.1 ± 3.4%, mean ± s.e.m.) similar to that of Esr1 expression (43.5 ± 2.5%; Fig. 1k–y). Double-labelling experiments confirmed a high degree of overlap (~90%) between recombined marker$^+$ and Esr1$^+$ cells in VMHvl (Fig. 1v–y), without spillover into the arcuate nucleus (Extended Data Fig. 1e–g). To activate Esr1$^+$ neurons optogenetically, *Esr1*$^{cre/+}$ male mice were unilaterally injected in VMHvl with an AAV encoding a Cre-dependent channelrhodopsin 2 (ref. 8) and a nuclear humanized *Renilla* green fluorescent protein (hrGFP) reporter (Fig. 2a). Photostimulation-dependent activation of Esr1$^+$ neurons was confirmed *in vitro* using whole-cell patch-clamp recording in acute hypothalamic slices (Fig. 2b–d), and *in vivo* by double-labelling for hrGFP and c-Fos (Fig. 2e–k), as well as by extracellular recordings (Extended Data Fig. 2).

Using an implanted fibre optic cable[9], we tested the effect of optogenetic stimulation of VMHvl Esr1$^+$ neurons in resident males in their home cage under infrared light, using the resident–intruder assay[10]. Stimulation (20 Hz, 30 s, 20 ms pulse-width) elicited intense, time-locked attack towards both castrated male and female intruders (Fig. 2l, m), in over 87% of ChR2-expressing animals and in ~90% of trials in those animals (Fig. 2n, o). Controls expressing Cre-dependent mCherry virus in VMHvl failed to show aggression during photostimulation (Fig. 2n, '0'; Extended Data Fig. 3a). Attack was initiated within ~5 s of photostimulation when light pulses were delivered while the resident was facing the intruder and within one mouse body-length (Extended Data Fig. 4),

and continued through most of the 30-s stimulation period (Fig. 2p, q; Supplementary Video 1). Optogenetic stimulation of VMHvl Esr1$^+$ neurons in females induced social investigation and occasional mounting, but not attack (Extended Data Fig. 5), indicating that sex differences in aggression probably occur within or downstream of VMHvl (refs 7, 11).

To determine whether non-Esr1-expressing VMHvl neurons contribute to aggression, we injected an AAV in which Cre recombination excises the ChR2–EYFP (enhanced yellow fluorescent protein) coding sequence (Fig. 2r, 'Cre-out'). Photostimulation failed to elicit any attack behaviour in these mice, but did elicit attack behaviour in wild-type mice injected with the same virus (Fig. 2s and Extended Data Fig. 3b, c). Together, these data indicate that optogenetic activation of VMHvl Esr1$^+$ neurons, but not of Esr1$^-$ neurons, is sufficient and specific for attack.

Previous loss-of-function manipulations in VMHvl, including GluCl-mediated neuronal silencing[3], ablation of progesterone receptor$^+$ neurons[11] and RNA interference-mediated knockdown of *Esr1* mRNA[12], reduced aggression, but such manipulations required a time scale of days or weeks. Therefore they did not distinguish whether these neurons are required to sense conspecifics, for actual attack, or for both. To distinguish these possibilities, we performed time-resolved, reversible optogenetic inhibition of VMHvl Esr1$^+$ neurons using eNpHR3.0[13]. Whole-cell patch-clamp recordings confirmed efficient photostimulation-dependent (532 nm) silencing of Esr1$^+$ neurons (Fig. 2u). Bilateral silencing (10 s continuous illumination) during an agonistic encounter interrupted attack in <3 s in ~60% of stimulation trials, with a median attack duration of ~2 s (Fig. 2v–y). In some trials, ongoing attack was abrogated almost instantaneously by photostimulation (Supplementary Video 2). Photostimulation also prevented the initiation of attack, and sometimes caused retreat from the intruder (Supplementary Video 3), when delivered at the moment of approach (Fig. 2z). *Esr1*$^{cre/+}$ males expressing mCherry showed no interruption of attack (Fig. 2v–y). Thus Esr1$^+$ neuronal activity is required for both the initiation and continuation of attack.

At early stages of a social encounter, resident males exhibit close investigation of intruders, with sniffing of the ano-genital and head regions (Fig. 3a)[14]. Under conditions of weak ChR2 expression or low-intensity photostimulation, when attack was usually not evoked (see Fig. 4 and Extended Data Fig. 6), we observed an increase in both the average number and duration of close investigation episodes during 30 s photostimulation trials, irrespective of the sex of the intruder (Fig. 3c, d, solid vs open red bars). This phenotype was observed in females as well as males (Extended Data Fig. 5 and Supplementary Video 4). We also observed a more aggressive form of close investigation during photostimulation, in which the resident vigorously pushed his nose into the intruder's anogenital region, in ~25% of mice (Fig. 3e and Supplementary Video 5). Importantly, bilateral optogenetic inhibition using eNpHR3.0 interrupted ongoing close investigation in ~60% of resident males, vs <20% of controls (Fig. 3f). Thus, VMHvl Esr1$^+$ neurons are necessary and sufficient for the investigative phase of a social encounter, as well as for attack.

Surprisingly, optogenetic stimulation under such conditions also promoted mounting behaviour, towards intact and castrated males as well as females (Fig. 3a, green rasters; Supplementary Videos 6 and 7, Extended Data Fig. 6) in ~50% of ChR2-expressing resident males and in ~60% of photostimulation trials for such males (Fig. 3g, h). Photostimulation of ChR2-expressing residents increased both the total number of mounts and average duration of mounting, towards both males and females (Fig. 3i, j). The frequency of evoked mounting was similar to that of control resident males towards female intruders (Fig. 3g, ♀, open bar). However when directed towards males, it was typically abortive and did not proceed to pelvic thrusting or ejaculation. Whereas male-directed mounting was only observed during photostimulation (Fig. 3a), its latency (~8–12 s; Fig. 3k) was longer than that for attack (~5 s). Photostimulation-induced mounting towards male intruders was not observed in mCherry-expressing controls (Fig. 3g, ♂, '0'). In

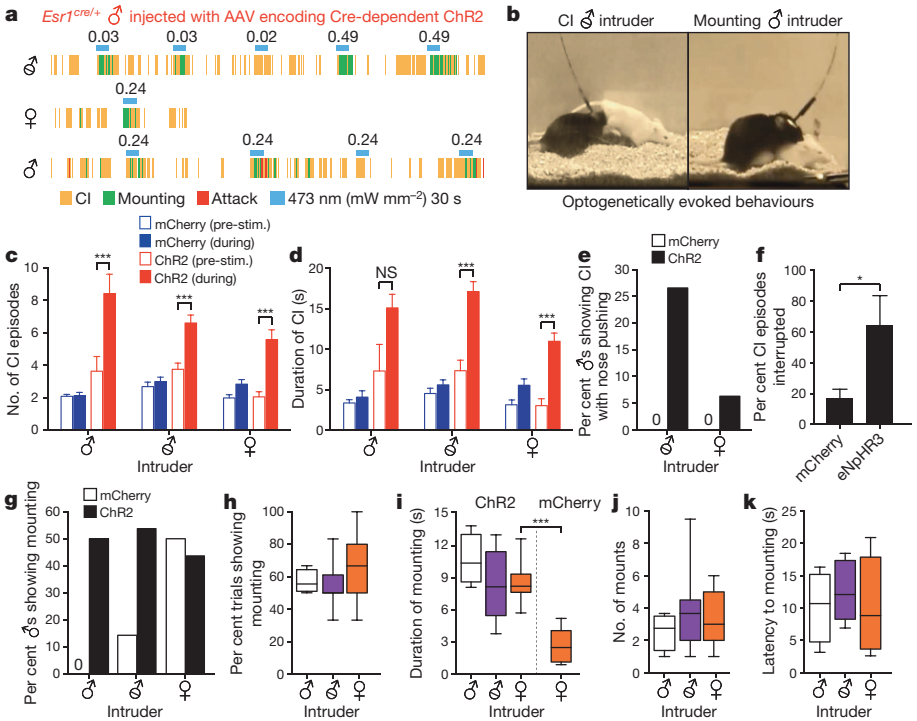


Figure 3 | Esr1$^+$ cells in VMHvl mediate close investigation (CI) and mounting behaviours. **a**, **b**, Representative raster plots (**a**) and video stills (**b**) illustrating photostimulation-evoked mounting (**a**, green; **b**, right) or CI (**a**, yellow; **b**, left) in ChR2-expressing *Esr1*$^{cre/+}$ males towards intruders of the indicated sex. **c**, **d**, Number (**c**) and duration (**d**) of CI episodes performed by males expressing mCherry (blue bars) or ChR2 (red bars), before (open bars) or during (filled bars) photostimulation, towards intruder males ($n = 4$ each), castrated males ($n = 14$ and $n = 18$, respectively) or females ($n = 10$ and $n = 12$). ***$P < 0.001$; two-way ANOVA with Tukey's multiple comparisons test. **e**, Aggressive sniffing ('CI with nose pushing') during photostimulation by *Esr1*$^{cre/+}$ males expressing mCherry (open bars) or ChR2 (filled bars) towards intruder castrated males ($n = 14$ and $n = 49$, respectively) or females ($n = 12$ and $n = 32$). **f**, Percentage of CI episodes interrupted by photostimulation of *Esr1*$^{cre/+}$ males expressing mCherry ($n = 7$) or eNpHR3.0 ($n = 3$). *$P < 0.05$; Mann–Whitney U-test. **g**. Percentage of *Esr1*$^{cre/+}$ males expressing mCherry (open bars) or ChR2 (filled bars) showing photostimulation-evoked mounting towards intruder males ($n = 4$ and $n = 8$, respectively), castrated males ($n = 14$ and $n = 35$) or females ($n = 10$ and $n = 28$). **h**–**k**, Quantification of photostimulation-evoked mounting towards intruder males ($n = 4$), castrated males ($n = 11$) or intact females ($n = 11$; **i**, mCherry, $n = 5$) towards females. ***$P < 0.001$; Mann–Whitney U-test. Data are mean ± s.e.m. (**c**, **d**) or median ± minimum and maximum values (**h**–**k**).

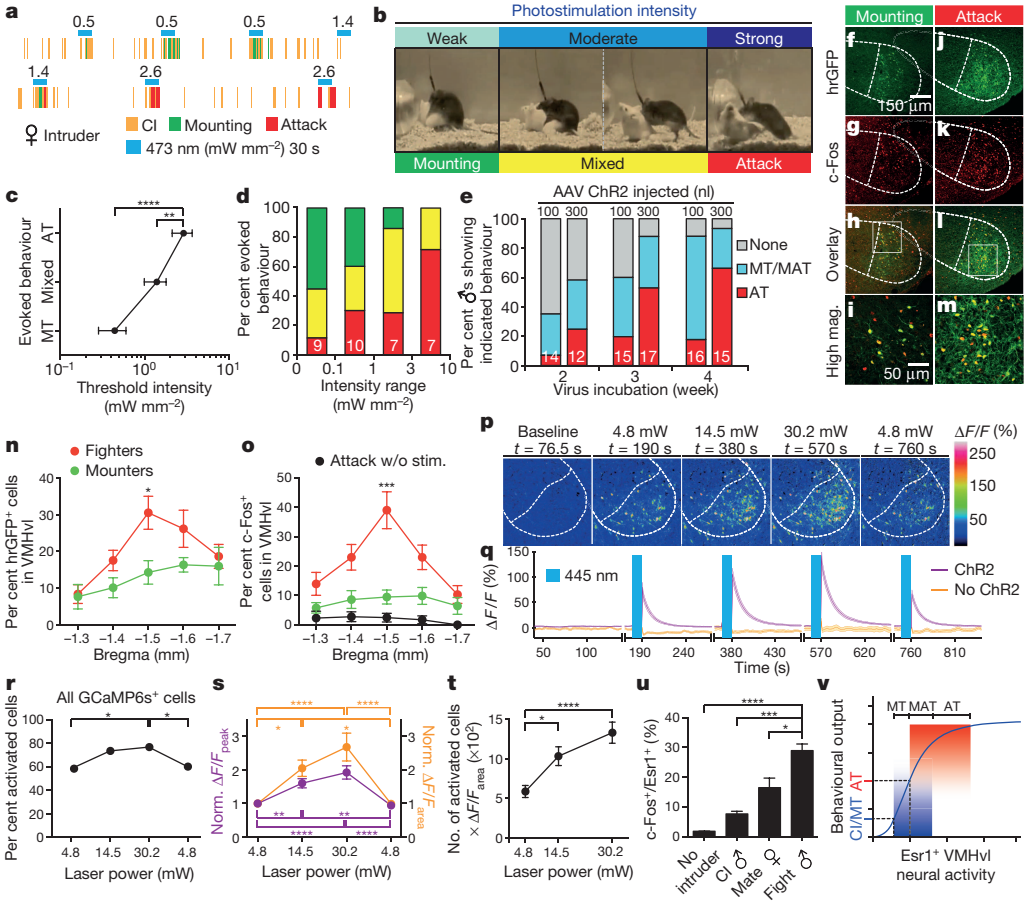


Figure 4 | Behavioural responses shift from mounting to attack depending on photostimulation intensity and the number of activated cells in VMHvl. **a**, **b**, Representative raster plots (**a**) and images (**b**) illustrating shift from mounting to attack with increasing photostimulation intensity. **c**, **d**, Threshold photostimulation intensities evoking mounting, mixed or attack behaviour (**c**, $n = 11$), or the relative percentage of those behaviours evoked by the indicated intensity range (**d**, n in bars). Data are from test sessions exhibiting multiple behaviours. $^{**}P < 0.01$, $^{****}P < 0.0001$; two-way ANOVA with Tukey's multiple comparisons test. **e**, Proportion of photostimulation-evoked behaviours in animals injected with different amounts of AAV (100 vs 300 nl), and incubated for different times post-injection (weeks). n within bars. Blue shading, mounting episodes with or without attack ('MT/MAT'). **f**–**m**, Double labelling for virally expressed hrGFP (**f**, **j**, **h**, **i**, **l**, **m**, native fluorescence) and photostimulation-induced c-Fos (**g**, **k**, **i**, **m**, anti-c-Fos, red) in solitary ChR2-expressing *Esr1*$^{cre/+}$ males photostimulated 1 h before euthanasia in their home cage, at an intensity that had previously evoked either mounting (**f**–**i**) or attack (**j**–**m**) several days earlier. **i**, **m**, Boxed areas in **h** and **l**, respectively. **n**, **o**. Quantification of number of hrGFP$^+$ (**n**) and c-Fos$^+$ (**o**) cells in VMHvl across successive axial levels, following solitary photostimulation of 'mounters' (green symbols, $n = 6$) or 'attackers' (red symbols, $n = 10$), as in **f**–**m**. Black symbols, no photostimulation before euthanasia ($n = 3$). $^{*}P < 0.05$, $^{***}P < 0.001$; two-way ANOVA with Tukey's multiple comparisons test. **p**–**t**, Two-photon Ca^{2+} imaging of acute hypothalamic slices expressing Cre-dependent ChR2–EYFP and Cre-independent GCaMP6s. **p**, Representative fluorescence images at the indicated time points and illumination power. **q**, Average Ca^{2+} transients in GCaMP6s$^+$ cells with (purple trace, $n = 60$) or without ChR2 (orange trace, $n = 48$). Photostimulation (445 nm, 2-ms pulses, 20 Hz) was delivered for 10 s (blue bars in **q**). **r**, Percentage of GCaMP6s-expressing cells with $\Delta F/F_{peak} > 5$ standard deviations from baseline, as a function of light power. $^{*}P < 0.05$; Pearson's Chi-square test. **s**, Normalized $\Delta F/F_{peak}$ (purple) and $\Delta F/F_{area}$ (orange, integrated area under the curves in **q** during 30 s following photostimulation) as a function of light power, relative to cells activated at 4.8 mW ($n = 35$). **t**, Number of activated cells × integrated activity per active cell, vs laser power. **s**, **t**, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$; repeated measures one-way ANOVA. **u**, Percentage of c-Fos$^+$ cells among Esr1$^+$ neurons in VMHvl of wild-type animals following the indicated behaviours (control, $n = 3$; CI, $n = 4$; mate, $n = 5$; fight, $n = 5$). Data are mean ± s.e.m. n = number of animals (**c**–**e**, **n**, **o**, **u**) or cells (**q**–**s**). **v**, Threshold model for relationship between level of Esr1$^+$ neuron activity and behaviour. MT, mount only; MAT, mixed mount and attack; AT, attack only. See also Supplementary Note 2.

contrast, optogenetic inhibition did not interrupt ongoing male–female mounting, or reduce its duration or frequency (Extended Data Fig. 7 and Supplementary Note 1).

To characterize the conditions that evoked mounting vs attack, we manipulated the amount of virus injected; the incubation time; and the intensity and frequency of photostimulation (Fig. 4 and Extended Data Fig. 8). Mounting behaviour was typically observed after short (<3 weeks) incubation times following viral injection; when smaller amounts of virus were injected (100 vs 300 nl); or at lower stimulation intensities (Fig. 4a–e). In some animals, as the intensity of stimulation was increased during a single trial, evoked behaviours could be observed to switch from mounting only (Fig. 4a, b, d, green bars), to mixed mounting and attack (Fig. 4b, d, yellow bars) to attack only (Fig. 4a, b, d, red bars; Extended Data Fig. 6d and Supplementary Videos 8–10). This effect often exhibited hysteresis: once attack was elicited, reducing the photostimulation intensity did not evoke mounting, but simply failed to elicit attack (Extended Data Fig. 9). A plot of log (photostimulation intensity) vs behaviour yielded a linear relationship, with mounting (either alone or mixed with attack) evoked up to 1.4 mW mm^{-2} and attack (only) evoked by intensities >3 mW mm^{-2} (Fig. 4c). Optogenetically evoked mounting vs attack behaviours were not strongly dependent on the frequency of photostimulation, although a higher percentage of stimulation trials elicited attack at 20 Hz than at 10 Hz (Extended Data Fig. 8).

The observation that the frequency of optogenetically evoked attack was higher in animals injected with larger volumes of virus (Fig. 4e, 300 nl) indicated that attack required activation of a larger number of

Esr1$^+$ neurons. Indeed, mice in which photostimulation induced attack contained a significantly higher number of virally infected (hrGFP$^+$) cells in VMHvl, than did animals in which it evoked mounting (Fig. 4f, j, n). Moreover, photostimulation of solitary residents 1 h before euthanasia yielded a significantly higher number of c-Fos$^+$ cells within the ChR2-expressing (hrGFP$^+$) population in 'attackers' (80.5 ± 4.9%, mean ± s.e.m.) than in 'mounters' (54.7 ± 12.6%, $P < 0.05$, two-tailed Student's *t*-test; Fig. 4f–m, o). These data indicate that optogenetically evoked attack requires a larger number of ChR2-expressing and active Esr1$^+$ cells, than does mounting. Consistent with this conclusion, the percentage of c-Fos$^+$ cells among Esr1$^+$ neurons in wild-type, unmanipulated animals was significantly higher following naturally occurring fighting, than after mating or close investigation (Fig. 4u)[3].

To investigate within the same preparation the relationship between the intensity of photostimulation and the number of active neurons in VMHvl, we used calcium imaging with a genetically encoded calcium indicator, GCaMP6s[15], to measure the extent of optogenetic activation in acute hypothalamic slices from *Esr1*$^{cre/+}$ males expressing ChR2 (Fig. 4p, q). Increasing the laser power over an approximately sixfold range significantly increased the fraction of GCaMP6s$^+$ cells exhibiting photostimulation-induced calcium transients (Fig. 4r), and also increased the peak and average activity per cell among neurons activated at the lowest power tested (Fig. 4q, s). There was a roughly linear relationship between laser power and overall activity (number of active neurons × average activity/neuron), with a slope of about 30 (Fig. 4t). No response was observed in controls lacking ChR2 (Fig. 4q, orange traces). Thus increasing light intensity augmented both the number of active neurons, and the average level of activity per neuron.

Earlier studies revealed that VMHvl contains neurons activated during male aggression, whose optogenetic stimulation or pharmacogenetic inhibition[16] evoked or inhibited attack, respectively[3]. However, those studies used ubiquitous promoters, and therefore did not identify the subpopulation of neurons responsible for attack. Here we identify these neurons as a subset (~40%) of VMHvl cells expressing Esr1. From one perspective, it is incidental that this marker encodes a hormone receptor. Indeed, our knock-in mice were designed to permit functional manipulations of these neurons without perturbing Esr1 function. Nevertheless, numerous genetic and pharmacologic studies have demonstrated steroid hormonal influences on the developmental and adult control of social behaviours, including those exerted via VMH[12,17–20] (reviewed in refs 21–23). However, relatively little is known about the circuit-level function of the neurons that express these receptors[24]. Genetic ablation of a closely related population of cells expressing the progesterone receptor partially reduced both male mating and aggression[11], but such loss-of-function manipulations do not have the temporal resolution of optogenetics. To our knowledge, therefore, the present experiments are the first to report time-resolved gain- and loss-of-function manipulations of hypothalamic neurons that express a sex-steroid hormone receptor. They reveal a complex and dynamic relationship between neuronal activity in this population and social behaviour. The relationship of this activity to hormonal influences remains to be investigated.

Previous *in vivo* recordings and c-Fos analysis revealed that VMHvl also contains neurons activated during male–female encounters; however, no effects of functional perturbations on male mating behaviour were observed[3] (see Supplementary Note 1). Those studies were, therefore, compatible with the view that VMHvl contains 'command'-like neurons[25] that control attack[26], and that the female-activated neurons might even serve to inhibit such attack neurons during mating[27]. The experiments reported here suggest a rather different view of VMHvl function. They show that Esr1$^+$ neurons control different behaviours throughout the entire progression of a social interaction, from its appetitive through its consummatory phases[1]. The fact that these behaviours are evoked by optogenetic activation at low and high photostimulation intensities, respectively, further suggests that increasing activity in VMHvl, as the intensity of the social encounter escalates[3], leads to qualitatively different behavioural outputs (Fig. 4v).

Several models may explain how such an apparent intensity coding of social behaviour is implemented at the cellular level (Supplementary Note 2). These include different subpopulations of Esr1$^+$ neurons with different activation thresholds, graded changes in activity within a single population (Extended Data Fig. 10), or more complex models involving attractor dynamics[28]. Although further experiments will be required to distinguish these possibilities, *in vivo* recordings revealed a substantial degree of overlap between neurons activated during mating vs fighting[3]. Whatever the explanation, the data suggest that Esr1$^+$ neurons in VMHvl may comprise a node in which a graded variable, perhaps representing the level of social arousal or cumulative sensory input, is transformed into differences in action selection at different thresholds. How this transformation occurs will be an interesting subject for future study.

Note added in proof: A detailed quantitative analysis of VMHvl spiking activity during male–male social interactions (Falkner *et al.*, *J. Neurosci.*, in the press) reveals that many individual neurons are active during both social investigation and attack; that spiking activity ramps up during the transition from investigation to attack; and that higher spiking activity during investigation predicts an increased likelihood of transitioning to attack. These results are consistent with the optogenetic experiments reported here.

METHODS SUMMARY

Generation of *Esr1*^{cre} knock-in mice. *Esr1*^{cre} knock-in mice were generated from 129S6/EvSvTac embryonic stem cells by targeting the 3′ end of the *Esr1* coding sequence with a cassette consisting of an *F2A* sequence, a *Cre* recombinase coding sequence, and an *Frt*-flanked PGK-neomycin resistance as an in-frame fusion. Chimaeric animals were backcrossed ($n > 6$) to C57BL/6N. The *Frt*-flanked PGK-neo-poly A cassette was removed by crossing *Esr1*$^{cre/+}$ mice to transgenic mice expressing flippase (FLPo, Jackson Laboratory, Stock No. 011065). The *Esr1*^{cre} knock-in mice are available from the Jackson Laboratory (Stock No. 017911).

Optogenetic activation and silencing of Esr1$^+$ neurons in VMHvl. Adult *Esr1*$^{cre/+}$ male and female mice at ages between 2 to 4 months were stereotaxically injected into the VMHvl with an AAV encoding Cre-dependent effector molecule (ChR2 or eNpHR3.0) or control fluorescent protein (mCherry, tdTomato or EGFP) following the procedures described previously[3]. After a 2–4-week recovery period, the virus-injected animals were examined behaviourally in their home cage using the resident–intruder assay for the next 2 to 5 weeks (1–2 test sessions per week). Photostimulation (ChR2, 473 nm, 20-ms pulses; eNpHR3.0, 593 nm, continuous) was applied to the virus-injected *Esr1*$^{cre/+}$ mice during social encounters with intruders (intact male, castrated male, or intact non-hormone primed female) 5–15 times per intruder.

All animal experiments were carried out in accordance with NIH guidelines and approved by the Institutional Animal Care and Use Committee (IACUC) at Caltech.

Electrophysiological recordings and calcium imaging. Whole-cell current clamp recordings and GCaMP6s-mediated two-photon calcium imaging were performed on acute hypothalamic slices prepared from virus-injected adult *Esr1*$^{cre/+}$ mice following the protocols described previously[3,29,30]. *In vivo* electrophysiology was performed as described[3], using a modification of the 16-wire electrode bundle with an integrated optic fibre.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 September 2013; accepted 21 February 2014.
Published online 16 April 2014.

1. Tinbergen, N. in *Physiological Mechanisms in Animal Behaviour* Vol. IV *Symposia of the Society for Experimental Biology* 305–312 (Academic Press, 1950).
2. Devidze, N., Lee, A., Zhou, J. & Pfaff, D. CNS arousal mechanisms bearing on sex and other biologically regulated behaviors. *Physiol. Behav.* **88,** 283–293 (2006).
3. Lin, D. *et al.* Functional identification of an aggression locus in the mouse hypothalamus. *Nature* **470,** 221–226 (2011).
4. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445,** 168–176 (2007).
5. Morgan, J. I., Cohen, D. R., Hempstead, J. L. & Curran, T. Mapping patterns of c-fos expression in the central nervous system after seizure. *Science* **237,** 192–197 (1987).
6. Walter, P. *et al.* Cloning of the human estrogen receptor cDNA. *Proc. Natl Acad. Sci. USA* **82,** 7889–7893 (1985).

7. Xu, X. *et al.* Modular genetic control of sexually dimorphic behaviors. *Cell* **148,** 596–607 (2012).
8. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neurosci.* **8,** 1263–1268 (2005).
9. Aravanis, A. M. *et al.* An optical neural interface: *in vivo* control of rodent motor cortex with integrated fiberoptic and optogenetic technology. *J. Neural Eng.* **4,** S143–S156 (2007).
10. Blanchard, D. C. & Blanchard, R. J. Ethoexperimental approaches to the biology of emotion. *Annu. Rev. Psychol.* **39,** 43–68 (1988).
11. Yang, C. F. *et al.* Sexually dimorphic neurons in the ventromedial hypothalamus govern mating in both sexes and aggression in males. *Cell* **153,** 896–909 (2013).
12. Sano, K., Tsuda, M. C., Musatov, S., Sakamoto, T. & Ogawa, S. Differential effects of site-specific knockdown of estrogen receptor alpha in the medial amygdala, medial pre-optic area, and ventromedial nucleus of the hypothalamus on sexual and aggressive behavior of male mice. *Eur. J. Neurosci.* **37,** 1308–1319 (2013).
13. Gradinaru, V. *et al.* Molecular and cellular approaches for diversifying and extending optogenetics. *Cell* **141,** 154–165 (2010).
14. Blanchard, R. J., Wall, P. M. & Blanchard, D. C. Problems in the study of rodent aggression. *Horm. Behav.* **44,** 161–170 (2003).
15. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).
16. Lerchner, W. *et al.* Reversible silencing of neuronal excitability in behaving mice by a genetically targeted, ivermectin-gated Cl^- channel. *Neuron* **54,** 35–49 (2007).
17. Cohen, R. S. & Pfaff, D. W. Ventromedial hypothalamic neurons in the mediation of long-lasting effects of estrogen on lordosis behavior. *Prog. Neurobiol.* **38,** 423–453 (1992).
18. Musatov, S., Chen, W., Pfaff, D. W., Kaplitt, M. G. & Ogawa, S. RNAi-mediated silencing of estrogen receptor α in the ventromedial nucleus of hypothalamus abolishes female sexual behaviors. *Proc. Natl Acad. Sci. USA* **103,** 10456–10460 (2006).
19. Spiteri, T. *et al.* Estrogen-induced sexual incentive motivation, proceptivity and receptivity depend on a functional estrogen receptor alpha in the ventromedial nucleus of the hypothalamus but not in the amygdala. *Neuroendocrinology* **91,** 142–154 (2010).
20. Spiteri, T. *et al.* The role of the estrogen receptor alpha in the medial amygdala and ventromedial nucleus of the hypothalamus in social recognition, anxiety and aggression. *Behav. Brain Res.* **210,** 211–220 (2010).
21. Simerly, R. B. Wired for reproduction: organization and development of sexually dimorphic circuits in the mammalian forebrain. *Annu. Rev. Neurosci.* **25,** 507–536 (2002).
22. Morris, J. A., Jordan, C. L. & Breedlove, S. M. Sexual differentiation of the vertebrate nervous system. *Nature Neurosci.* **7,** 1034–1039 (2004).
23. Wu, M. V. & Shah, N. M. Control of masculinization of the brain and behavior. *Curr. Opin. Neurobiol.* **21,** 116–123 (2011).
24. Kow, L. M., Easton, A. & Pfaff, D. W. Acute estrogen potentiates excitatory responses of neurons in rat hypothalamic ventromedial nucleus. *Brain Res.* **10,** 124–131 (2005).
25. Bentley, D. & Konishi, M. Neural control of behavior. *Annu. Rev. Neurosci.* **1,** 35–59 (1978).
26. Kruk, M. R. *et al.* Discriminant analysis of the localization of aggression-inducing electrode placements in the hypothalamus of male rats. *Brain Res.* **260,** 61–79 (1983).
27. Anderson, D. J. Optogenetics, sex, and violence in the brain: implications for psychiatry. *Biol. Psychiatry* **71,** 1081–1089 (2011).
28. Kristan, W. B. Neuronal decision-making circuits. *Curr. Biol.* **18,** R928–R932 (2008).
29. Haubensak, W. *et al.* Genetic dissection of an amygdala microcircuit that gates conditioned fear. *Nature* **468,** 270–276 (2010).
30. Vrontou, S., Wong, A. M., Rau, K. K., Koerber, H. R. & Anderson, D. J. Genetic identification of C fibres that detect massage-like stroking of hairy skin *in vivo*. *Nature* **493,** 669–673 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Park for behavioural scoring, R. Robertson for behavioural scoring and MATLAB programming, L. Lo for testing Cre-mediated recombination in *Esr1*$^{cre/+}$ male mice, C. Chiu and X. Wang for histology, M. McCardle for genotyping, J. S. Chang for technical assistance, S. Pease for generation of knock-in mice, H. Cai for training in slice electrophysiology, A. Wong for assistance with two-photon imaging, K. Deisseroth and J. Harris for AAV constructs, E. Boyden for advice on ferrule fibre fabrication, D. Lin and M. Boyle for their contributions to early stages of this project, W. Hong and R. Axel for comments on the manuscript, C. Chiu for laboratory management and G. Mancuso for administrative assistance. D.J.A. is an Investigator of the Howard Hughes Medical Institute and a Paul G. Allen Distinguished Investigator. This work was supported in part by NIH grant no. R01MH070053, and grants from the Gordon Moore Foundation and Ellison Medical Research Foundation. H.L. was supported by the NIH Pathway to Independence Award 1K99NS074077. T.E.A. was supported by NIH NRSA postdoctoral fellowship grant 1F32HD055198-01 and a Beckman Fellowship.

Author Contributions H.L. characterized *Esr1*^{cre} mice, designed and performed optogenetic behavioural experiments and co-wrote the manuscript; D.-W.K. performed slice electrophysiology and imaging experiments; R.R. performed *in vivo* electrophysiology; T.E.A. generated the *Esr1*^{cre} targeting construct and AAV vectors; A.C. carried out some behavioural experiments; L.M. and H.Z. performed *in situ* hybridization experiments; D.J.A. supervised experiments and co-wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.J.A. (wuwei@caltech.edu).

LETTER

doi:10.1038/nature13315

A single female-specific piRNA is the primary determiner of sex in the silkworm

Takashi Kiuchi[1], Hikaru Koga[1]*, Munetaka Kawamoto[1]*, Keisuke Shoji[1]*, Hiroki Sakai[2], Yuji Arai[1], Genki Ishihara[1], Shinpei Kawaoka[1], Sumio Sugano[3], Toru Shimada[1], Yutaka Suzuki[3], Masataka G. Suzuki[2] & Susumu Katsuma[1]

The silkworm *Bombyx mori* uses a WZ sex determination system that is analogous to the one found in birds and some reptiles. In this system, males have two Z sex chromosomes, whereas females have Z and W sex chromosomes. The silkworm W chromosome has a dominant role in female determination[1,2], suggesting the existence of a dominant feminizing gene in this chromosome. However, the W chromosome is almost fully occupied by transposable element sequences[3–5], and no functional protein-coding gene has been identified so far. Female-enriched PIWI-interacting RNAs (piRNAs) are the only known transcripts that are produced from the sex-determining region of the W chromosome[6], but the function(s) of these piRNAs are unknown. Here we show that a W-chromosome-derived, female-specific piRNA is the feminizing factor of *B. mori*. This piRNA is produced from a piRNA precursor which we named *Fem*. *Fem* sequences were arranged in tandem in the sex-determining region of the W chromosome. Inhibition of *Fem*-derived piRNA-mediated signalling in female embryos led to the production of the male-specific splice variants of *B. mori doublesex* (*Bmdsx*), a gene which acts at the downstream end of the sex differentiation cascade[7,8]. A target gene of *Fem*-derived piRNA was identified on the Z chromosome of *B. mori*. This gene, which we named *Masc*, encoded a CCCH-type zinc finger protein. We show that the silencing of *Masc* messenger RNA by *Fem* piRNA is required for the production of female-specific isoforms of *Bmdsx* in female embryos, and that Masc protein controls both dosage compensation and masculinization in male embryos. Our study characterizes a single small RNA that is responsible for primary sex determination in the WZ sex determination system.

In *Bombyx mori*, sex determination is probably established at an early stage of embryogenesis. We prepared the sexed RNA from individual silkworm embryos genotyped by three W chromosome-specific randomly amplified polymorphic DNA (RAPD) markers[4] (Extended Data Fig. 1a), and examined the splicing pattern of a *doublesex* orthologue of *B. mori* (*Bmdsx*). *Bmdsx* produces female- and male-specific RNAs by sex-specific alternative splicing[9] that have essential roles in silkworm sexual development[7,8]. Female-specific splice variants of *Bmdsx* were the default transcripts during an early stage of development. Whereas the male-specific splice variants clearly appeared in male embryos, only faint bands were observed in females, from 21–24 h post-oviposition (hpo) (Fig. 1a). This indicated that the feminizing signal is transmitted from the W chromosome before 21 hpo. Thus, we performed deep sequencing of RNAs (RNA-seq) isolated from male and female embryos at 15, 18, 21 and 24 hpo, and identified differentially expressed transcripts between male and female embryos.

One contig, comp73859_c0, was consistently identified in female embryos at all of the developmental times tested (Extended Data Fig. 1b, c). This sequence was amplified by PCR only when female genomic DNA or complementary DNA was used as a template (Fig. 1b). The sequence of comp73859_c0 did not show significant identity with any sequence in the draft male silkworm genome sequence[10], suggesting that this contig is localized on and transcribed from the W chromosome. In addition, this sequence was amplified from genomic DNA isolated from female wild silkmoth *Bombyx mandarina*, but not from males (Fig. 1b). Reverse transcription followed by quantitative PCR (RT–qPCR) showed that the expression level peaked at 18–21 hpo, and then gradually declined during embryogenesis (Fig. 1c). This transcript was also detected in the ovary and other somatic tissues (Extended Data Fig. 2a, b, d). Long PCR demonstrated that there are multiple copies of this sequence on the W chromosome (Extended Data Fig. 2c). The sequences were occasionally arranged in tandem and some were probably expressed as long transcriptional units. Northern blot analysis revealed that transcripts of approximately 0.8 and 1.4 kilobase are major units, and antisense transcripts were not detected (Extended Data Fig. 2d). The copy number of this contig per haploid genome was estimated at more than 30 in the genome of *B. mori*.

This contig did not show homology with any known sequence, nor did it seem to encode a functional protein. Instead, this transcript seemed to be a piRNA precursor. Mapping of embryonic or ovarian piRNAs[6,11,12] onto this transcript and northern blotting revealed a 29-nucleotide-long piRNA-producing region (Fig. 1d and Extended Data Fig. 3a). This piRNA was poorly transmitted from the mother moth, accumulated from 15 hpo,

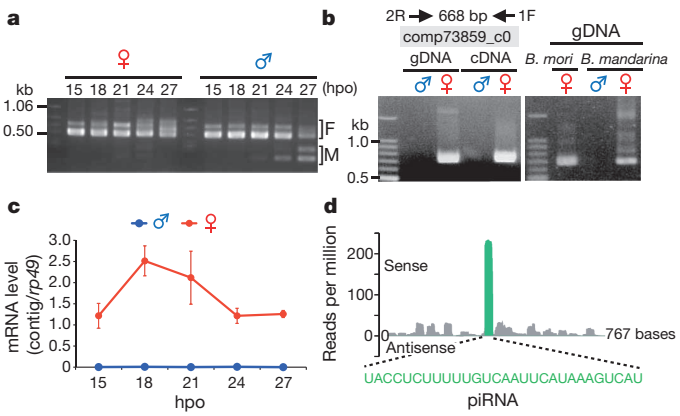


Figure 1 | Characterization of a female-specific piRNA precursor in early silkworm embryos. **a**, Splicing patterns of *Bmdsx* in early embryos. The F and M indicate female- and male-type splicing of *Bmdsx*, respectively. Similar results were obtained in three independent experiments. **b**, Detection of a W chromosome-derived transcript. Genomic DNA (gDNA) and cDNA were prepared from female and male embryos of *B. mori* at 24 hpo (left panel) or adult *B. mandarina* (right panel). **c**, Expression profile of the female-specific contig in early embryos. Data shown are means ± s.d. of three embryos. **d**, Mapping of embryonic piRNAs (24 hpo) onto the comp73859_c0. The relative location, abundance and sequence of the 29 base-long piRNA (shown in green) are indicated.

[1]Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan. [2]Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan. [3]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.
*These authors contributed equally to this work.

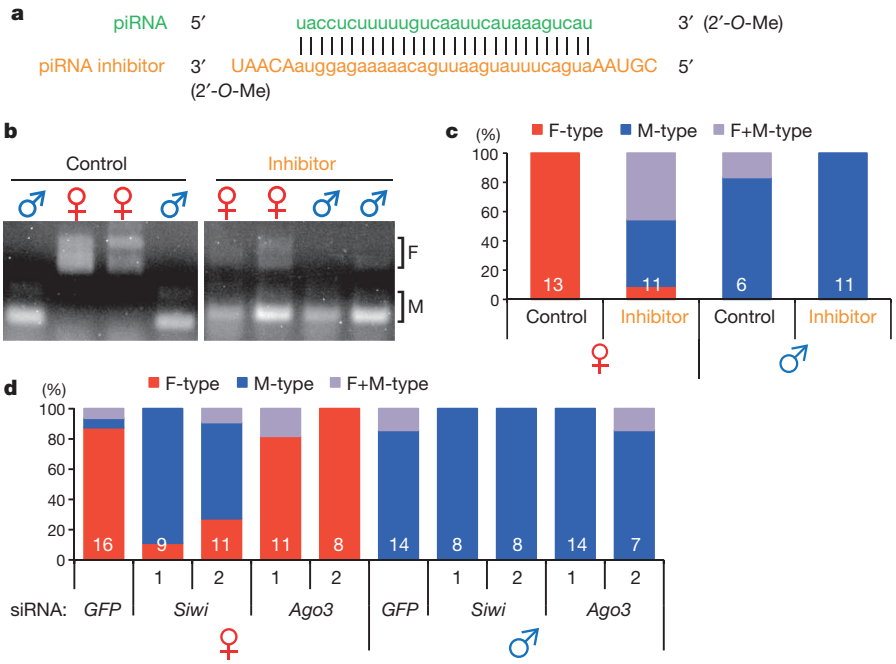


Figure 2 | Female-specific piRNA is a primary sex determinant of *B. mori*. **a**, Structure of the female-specific piRNA and its inhibitor. **b**, **c**, Effect of the RNA inhibitor on the *Bmdsx* splicing. The splicing patterns were examined at 72 h post-injection. Representative splicing patterns are shown in **b** and the data are summarized in **c**. The number indicates the sample size. F-type, female-type; M-type, male-type; F+M-type, both variants are mixed. **d**, Splicing of *Bmdsx* in embryos that were injected with *Siwi* or *BmAgo3* siRNAs. Two types of siRNAs for each target were used. The splicing patterns were examined at 72 h post-injection. The numbers in the columns indicate sample size.

and increased rapidly between 18 and 21 hpo (Extended Data Fig. 3b, c). By screening piRNA libraries that were generated from three *B. mori* strains that each possess a unique W chromosome structure[6], we found that this piRNA was produced from the sex-determining region of W chromosome (Extended Data Fig. 3d).

The silkworm KG strain, which possesses a mutation(s) in the W chromosome, shows various degrees of female masculinization features[13]. Expression of the contig-derived piRNA in the masculinized ovary was markedly lower than in wild type (Extended Data Fig. 4a), indicating that a certain amount of this piRNA might be required for complete feminization of adult moths. To investigate the role of this piRNA, we used a unique RNA-based inhibitor that seemed to function in BmN4 cells (Fig. 2a and Extended Data Fig. 4b) and investigated the effect of the inhibitor on *Bmdsx* splicing in early embryos. The *Bmdsx* splicing was markedly altered to produce the male-type isoform when the inhibitor was injected into female embryos, whereas the pattern was not affected in male embryos (Fig. 2b–c). The splicing pattern, however, was not altered in newly hatched larvae, presumably because this inhibitor does not possess the long-term inhibitory activity (Extended Data Fig. 4c). These results demonstrated that the targeted piRNA is required for the female-type splicing of *Bmdsx*. Thus, we named the precursor of this piRNA, *Feminizer* (*Fem*).

We performed RNA interference (RNAi) experiments that targeted two core components of the silkworm piRNA biogenesis pathway called *Siwi* and *BmAgo3* (Extended Data Fig. 4d)[14,15]. Small interfering RNA (siRNA)-mediated knockdown of *Siwi* in female embryos commonly led to the production of the male-type *Bmdsx* transcripts, whereas little effect was observed in male embryos (Fig. 2d and Extended Data Fig. 4e). *BmAgo3* RNAi did not affect *Bmdsx* splicing in either female or male embryos (Fig. 2d and Extended Data Fig. 4e). These results indicated that *Siwi* expression is crucial for the female-type splicing of *Bmdsx* in female embryos. Our hypothesis as to why *BmAgo3* knockdown did not affect *Bmdsx* splicing in early embryos is discussed later.

We identified only one genomic locus where the *Fem* piRNA sequence was extensively complementary (Fig. 3a). This locus was present within the ninth exon of an uncharacterized gene located on the Z chromosome (Fig. 3a and Extended Data Fig. 5a). We named this gene *Masculinizer* (*Masc*). *Masc* potentially encoded a novel CCCH-tandem zinc finger protein (Extended Data Fig. 5b). Phylogenetic analysis suggested that this protein forms a novel lepidopteran-specific protein family (Extended Data Fig. 5c). PIWI–piRNA complexes are known to cleave their complementary target sequences across from positions 10 and 11 of the guide piRNA[16,17]. By a modified 5′ rapid amplification of cDNA ends (modified RACE) method[18], we found that all of the cloned 5′ ends of the

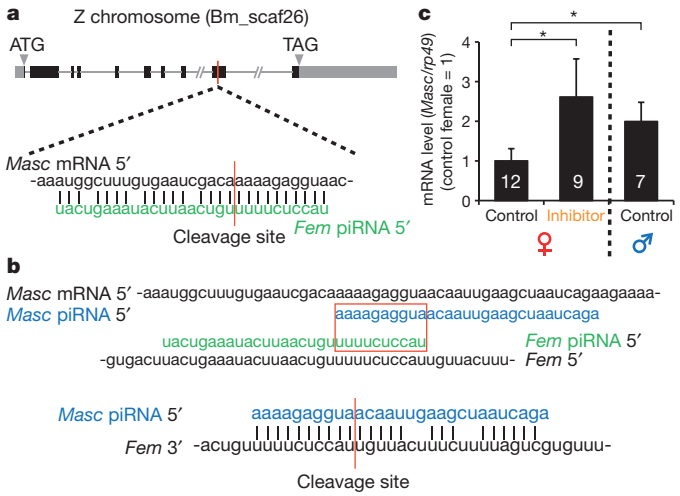


Figure 3 | *Masc* mRNA is the target of *Fem* piRNA. **a**, Genomic structure of the *Masc* gene on the Z chromosome of *B. mori*. The putative cleavage site by the *Fem* piRNA–Siwi complex is shown by the red line. **b**, Identification of *Masc* piRNA. The ping-pong signature (within the red box) and putative cleavage site (red line) of *Fem* by the *Masc* piRNA–BmAgo3 complex are shown. **c**, *Masc* expression in the piRNA inhibitor-injected embryos at 18 hpo. The numbers in the columns indicate sample size. Data shown are means + s.d. Data were subjected to Kruskal–Wallis analysis with post hoc Dunn's test. $*P < 0.05$.

Masc-derived RNA fragments from early embryos mapped precisely to the predicted *Fem* piRNA cleavage site (Extended Data Fig. 6a), indicating that *Masc* mRNA is the target of *Fem* piRNA.

The piRNA biogenesis occurs through a ping-pong mechanism that involves two different PIWI proteins. A 10-nucleotide overlap between sense and antisense piRNAs, called a ping-pong signature, is the hallmark of the cleavage reaction catalysed by PIWI proteins[16,17]. We found piRNAs that have a perfect 10-nucleotide overlap with *Fem* piRNA. The most abundant of these piRNAs were those that perfectly matched to the *Masc* coding region (Fig. 3b and Extended Data Fig. 6b, c), indicating that the *Masc* mRNA-derived piRNA (*Masc* piRNA) is a ping-pong partner of *Fem* piRNA. The *Masc* piRNA was extensively complementary to *Fem* (Fig. 3b), indicating that the PIWI–*Masc* piRNA complex will reliably slice *Fem* RNA. *Fem* piRNA preferentially bound to Siwi, whereas *Masc* piRNA preferentially bound to BmAgo3 (Extended Data Fig. 6d). Thus, our findings indicate a ping-pong amplification model for *Fem* and *Masc* piRNAs (Extended Data Fig. 6e). This model is experimentally supported by the introduction of the inhibitor for *Fem* piRNA (Fig. 3c) or siRNAs for *Siwi* (Extended Data Fig. 7a) into female embryos showing enhanced *Masc* levels. Unlike *Fem* piRNA, a moderate amount of *Masc* piRNA was maternally transmitted (Extended Data Fig. 6f, g). Together with the fact that BmAgo3 is also maternally transmitted[12], this suggests that a moderate amount of the *Masc* piRNA–BmAgo3 complex exists even in newly laid eggs. The presence of this complex helps to explain why *BmAgo3* RNAi in female embryos had little effect on *Bmdsx* splicing (Fig. 2d). Embryonic RNAi for *BmAgo3* did not alter the *Bmdsx* splicing, but enhanced *Masc* expression in newly hatched female larvae (Extended Data Fig. 7b–d), supporting the role of the *Masc* piRNA–BmAgo3 complex in sex determination. Higher levels of *Masc* piRNA were detected from 21–27 hpo (Extended Data Fig. 6f, g); this increase correlated with a massive accumulation of *Fem* piRNA (Extended Data Fig. 3b, c).

In male embryos, *Masc* expression rapidly increased, then rapidly decreased between 15 and 18, and 18 and 21 hpo, respectively (Fig. 4a). In contrast, *Masc* expression in female embryos gradually declined from 15 hpo, and remained at a low level compared with that found in males (Fig. 4a). These data indicate that *Fem* piRNA-mediated cleavage of *Masc* mRNA results in low-level accumulation of *Masc* mRNA in female embryos. Injection of *Masc* siRNA into male embryos reduced *Masc* expression to levels that were found in control female embryos at 18 hpo (Fig. 4b), and resulted in the production of female-type variants of *Bmdsx* throughout the embryonic stage (Fig. 4c and Extended Data Fig. 8a, b). Female embryos injected with *Masc* siRNA hatched normally, whereas male embryos did not (Fig. 4d), indicating that inhibition of the *Masc* pathway at the embryonic stage results in male-specific lethality. This probably mimics the way that an arthropod pathogen *Wolbachia* induces a male-killing phenotype in lepidopteran insects[19]. The *Fem* piRNA-resistant *Masc* (*Masc-R*) mRNA was more accumulated than the wild-type *Masc* mRNA in *Masc* cDNA-transfected BmN4 cells, whereas *Masc* piRNA was poorly detected in *Masc-R* cDNA-transfected cells (Extended Data Fig. 9a–c). The *Masc-R* mRNA was not cleaved, which changed the *Bmdsx* splicing pattern in BmN4 cells to the male-type completely, and induced a growth inhibition (Extended Data Fig. 9d–f), indicating that the *Fem* piRNA-mediated cleavage of *Masc* mRNA is essential for silkworm feminization.

RNA-seq analyses of *Masc* siRNA-injected embryos revealed that the transcripts differentially expressed in males were mapped predominantly onto the Z chromosome (chromosome 1, 51%), whereas such a bias was not observed in females (Fig. 4e). Most of the Z-chromosome-derived transcripts expressed differentially in males (97%) were expressed higher in *Masc* RNAi embryos (Fig. 4e) and randomly dispersed throughout this chromosome (Extended Data Fig. 10). These results demonstrate that Masc protein globally represses gene expression from the male Z chromosome at the embryonic stage. Taken together, Masc protein controls both dosage compensation and masculinization (Fig. 4f). In *Drosophila*, Sex-lethal, a master switch for sex determination, controls dosage

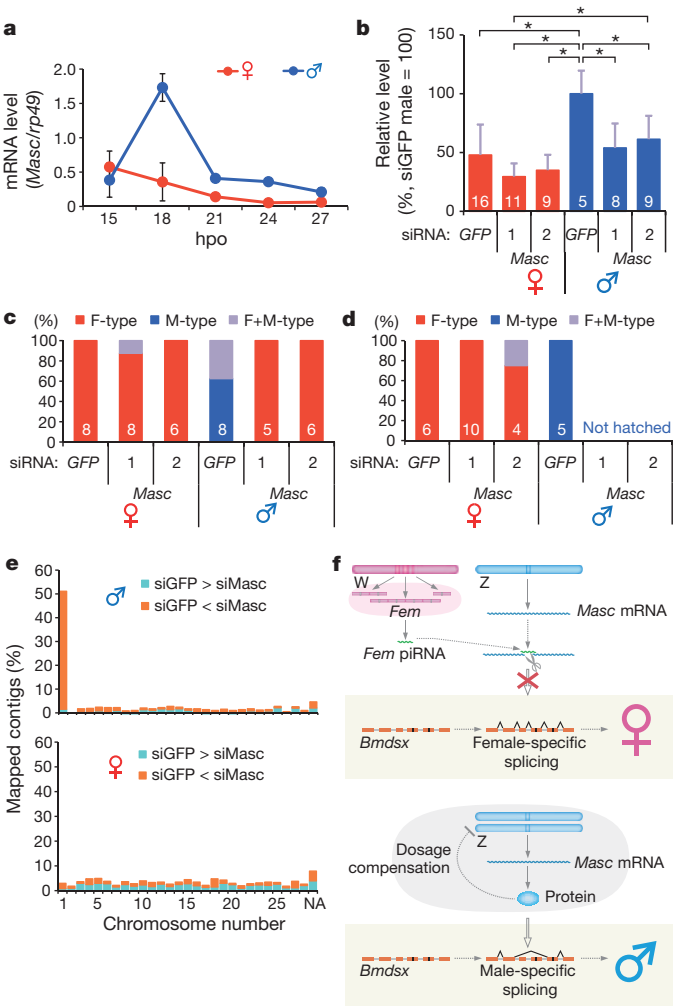


Figure 4 | Masc protein controls both masculinization and dosage compensation in male embryos. **a**, Expression profile of *Masc* in early embryos. Data shown are means ± s.d. of three embryos. **b**, Knockdown of *Masc* mRNA in *B. mori* embryos. The embryos were injected with two types of siRNAs for *Masc*, and *Masc* expression was examined by RT–qPCR at 18 hpo. Data shown are means + s.d. The number indicates the sample size. One-way ANOVA was performed with post hoc Tukey's test. $*P < 0.05$. **c**, **d**, Splicing of *Bmdsx* in *Masc* siRNA-injected embryos. The splicing pattern was determined at 72 h (**c**) and about 240 h (**d**, immediately after hatching) post-injection. The number indicates the sample size. **e**, Differentially expressed transcripts in *Masc* RNAi embryos. siGFP, *GFP* siRNA-injected embryos; siMasc, *Masc* siRNA-injected embryos; NA, not assigned. **f**, A proposed model for the sex determination pathway in *B. mori*.

compensation by inhibiting translation of *male-specific lethal 2* (*msl-2*)[20]. Loss of *msl-2* causes male lethality, owing to the failure of hypertranscription from the male X chromosome. A failure of dosage compensation is probably involved in male-specific lethality of *Masc* mRNA-depleted male embryos.

We unravelled a question that has perplexed insect geneticists for more than eight decades. Our study answers the question of how the W chromosome determines the femaleness of the silkworm *B. mori*. The silkworm feminizer *Fem* is the precursor of a 29-nucleotide-long small RNA. To our knowledge, this is the first example of the identification of a primary sex-determining factor in Lepidoptera, and the first experimental evidence showing a piRNA-mediated sex determination mechanism. Our findings also suggest that *Masc* levels may be involved in sex determination in lepidopteran species that are monosomic (Z0) in females and ZZ in males[21]. We are now experimentally surveying this hypothesis using moth species with a Z0/ZZ sex chromosome constitution.

METHODS SUMMARY

Sex-specific splicing of *Bmdsx*, piRNA mapping, qRT–PCR of piRNA, and transfection experiments using BmN4 cells were performed as described previously[6,12,13,14]. Embryonic RNAi was performed by injecting embryos with two different siRNAs for each gene investigated.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 October 2013; accepted 8 April 2014.
Published online 14 May 2014.

1. Hasimoto, H. The role of the W-chromosome in the sex determination of *Bombyx mori* [in Japanese]. *Jpn. J. Genet.* **8,** 245–247 (1933).
2. Tajima, Y. Studies on chromosome aberrations in the silkworm. II. Translocation involving second and W-chromosomes [in Japanese]. *Bull. Seric. Exp. Stn.* **12,** 109–181 (1944).
3. Abe, H. *et al.* Identification of novel random amplified polymorphic DNAs (RAPDs) on the W chromosome of the domesticated silkworm, *Bombyx mori*, and the wild silkworm, *B. mandarina*, and their retrotransposable element-related nucleotide sequences. *Genes Genet. Syst.* **73,** 243–254 (1998).
4. Abe, H. *et al.* Partial deletions of the W chromosome due to reciprocal translocation in the silkworm *Bombyx mori*. *Insect Mol. Biol.* **14,** 339–352 (2005).
5. Abe, H., Mita, K., Yasukochi, Y., Ohshiki, T. & Shimada, T. Retrotransposable elements on the W chromosome of the silkworm, *Bombyx mori*. *Cytogenet. Genome Res.* **110,** 144–151 (2005).
6. Kawaoka, S. *et al.* The silkworm W chromosome is a source of female-enriched piRNAs. *RNA* **17,** 2144–2151 (2011).
7. Suzuki, M. G., Funaguma, S., Kanda, T., Tamura, T. & Shimada, T. Analysis of the biological functions of a *doublesex* homologue in *Bombyx mori*. *Dev. Genes Evol.* **213,** 345–354 (2003).
8. Suzuki, M. G., Funaguma, S., Kanda, T., Tamura, T. & Shimada, T. Role of the male BmDSX protein in the sexual differentiation of *Bombyx mori*. *Evol. Dev.* **7,** 58–68 (2005).
9. Ohbayashi, F., Suzuki, M. G., Mita, K., Okano, K. & Shimada, T. A homologue of the *Drosophila doublesex* gene is transcribed into sex-specific mRNA isoforms in the silkworm, *Bombyx mori*. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **128,** 145–158 (2001).
10. The International Silkworm Genome Consortium. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* **38,** 1036–1045 (2008).
11. Kawaoka, S. *et al. Bombyx* small RNAs: genomic defense system against transposons in the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* **38,** 1058–1065 (2008).
12. Kawaoka, S. *et al.* Zygotic amplification of secondary piRNAs during silkworm embryogenesis. *RNA* **17,** 1401–1407 (2011).
13. Hara, K. *et al.* Altered expression of testis-specific genes, piRNAs, and transposons in the silkworm ovary masculinized by a W chromosome mutation. *BMC Genomics* **13,** 119 (2012).
14. Kawaoka, S. *et al.* The *Bombyx* ovary-derived cell line endogenously expresses PIWI/PIWI-interacting RNA complexes. *RNA* **15,** 1258–1264 (2009).
15. Kawaoka, S., Minami, K., Katsuma, S., Mita, K. & Shimada, T. Developmentally synchronized expression of two *Bombyx mori* Piwi subfamily genes, *SIWI* and *BmAGO3* in germ-line cells. *Biochem. Biophys. Res. Commun.* **367,** 755–760 (2008).
16. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128,** 1089–1103 (2007).
17. Gunawardane, L. S. *et al.* A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* **315,** 1587–1590 (2007).
18. Watanabe, T. *et al.* Role of piRNAs and noncoding RNA in de novo DNA methylation of the imprinted mouse *Rasgrf1* locus. *Science* **332,** 848–852 (2011).
19. Sugimoto, T. N. & Ishikawa, Y. A male-killing *Wolbachia* carries a feminizing factor and is associated with degradation of the sex-determining system of its host. *Biol. Lett.* **8,** 412–415 (2012).
20. Penalva, L. O. & Sánchez, L. RNA binding protein sex-lethal (Sxl) and control of *Drosophila* sex determination and dosage compensation. *Microbiol. Mol. Biol. Rev.* **67,** 343–359 (2003).
21. Traut, W., Sahara, K. & Marec, F. Sex chromosomes and sex determination in Lepidoptera. *Sex Dev.* **1,** 332–346 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. G. Kamita for critical reading of the manuscript; Y. Tomari for critical reading of the manuscript and technical suggestions. This work was supported by the Program for Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry to Su.K. and Grants-in-Aid for Scientific Research on Innovative Areas (Nos. 22115502 and 22128004) to Su.K. and T.S.

Author Contributions Su.K., T.K. and M.G.S. conceived and designed the experiments. T.K., H.K., K.S., H.S., G.I., Y.A., Sh.K., M.G.S. and Su.K. performed molecular biological experiments. M.K. and K.S. performed most of the bioinformatic analyses. S.S. and Y.S. performed deep sequencing and data analysis. T.S. provided essential reagents and expertise. All of the authors discussed the data and helped manuscript preparation. Su.K. wrote the manuscript with intellectual input from all authors. Su.K. supervised the project.

Author Information The nucleotide sequences of *Fem* and *Masc* have been deposited in the DDBJ/EMBL/GenBank data bank under the accession numbers AB840787 and AB840788. Deep sequencing data obtained in this study are available under the accession numbers DRA001104 and DRA001338 (DDBJ), respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Su.K. (katsuma@ss.ab.a.u-tokyo.ac.jp).

LETTER

doi:10.1038/nature13300

Clonal selection in the germinal centre by regulated proliferation and hypermutation

Alexander D. Gitlin[1], Ziv Shulman[1] & Michel C. Nussenzweig[1,2]

During immune responses, B lymphocytes clonally expand and undergo secondary diversification of their immunoglobulin genes in germinal centres (GCs)[1–4]. High-affinity B cells are expanded through iterative interzonal cycles of division and hypermutation in the GC dark zone followed by migration to the GC light zone, where they are selected on the basis of affinity to return to the dark zone[5–10]. Here we combine a transgenic strategy to measure cell division and a photoactivatable fluorescent reporter to examine whether the extent of clonal expansion and hypermutation are regulated during interzonal GC cycles. We find that both cell division and hypermutation are directly proportional to the amount of antigen captured and presented by GC B cells to follicular helper T cells in the light zone. Our data explain how GC B cells with the highest affinity for antigen are selectively expanded and diversified.

Clonal expansion is an essential feature of the immune response. B lymphocytes bearing antigen-specific immunoglobulins undergo this process in the germinal centre, a specialized microanatomical compartment where B cells also diversify their immunoglobulin genes through somatic hypermutation[1–4]. GC B cells expressing mutated surface immunoglobulins with the highest affinity are then positively selected by iterative cycles of cell division, somatic hypermutation and selection[5–10], endowing the host with high-affinity humoral immunity[4].

GC B cells divide and mutate in the dark zone (DZ), and then migrate to the light zone (LZ) where they capture antigen through surface immunoglobulin and present it as peptide bound to major histocompatibility complex class II (pMHCII) to cognate follicular helper T cells (T_{FH}) [4,10–12]. Migration between the two zones is mediated by the chemokine receptors CXCR4 and CXCR5, with 50% of DZ cells migrating to the LZ, and 10% returning to the DZ from the LZ within 6 h[5,10,13]. Moreover, B cells in the two GC zones alternate between distinct genetic programs reflecting cell division in the DZ and selection in the LZ, but do so independently of local cues received in the two zones[10,14]. However, the precise mechanism by which the highest affinity cells are selected, and whether cell divisions and immunoglobulin mutations in the DZ are regulated, remains unknown[14].

To determine whether the amount of antigen internalized by GC B cells governs the extent of clonal expansion, we titrated the amount of antigen delivered to GC B cells using antibodies that target DEC205, an endocytic receptor that carries antigen to intracellular MHCII-containing compartments[10,15–18]. GC responses were initiated by priming mice with ovalbumin (OVA), followed by boosting with OVA coupled to the hapten 4-hydroxy-3-nitrophenylacetyl (NP–OVA)[9]. Antigen-specific B-cell responses were tracked by adoptive transfer of B1-8^{hi} immunoglobulin heavy chain knock-in B cells, which are specific for NP when they express immunoglobulin lambda (Igλ) light chains[19]. To measure the relative expansion of B cells receiving graded amounts of antigen, GCs were induced in mice that received a mixture of B1-8^{hi} DEC205$^{+/+}$ and B1-8^{hi} DEC205$^{-/-}$ B cells at a 5:95 ratio. Graded doses of antigen were delivered to DEC205$^{+/+}$ GC B cells using the chimaeric DEC205 antibody fused to cognate antigen, OVA (αDEC–OVA, Fig. 1a)[20]. Whereas control injections with PBS had no effect, injection with 10 μg of αDEC–OVA resulted in selective expansion of the B1-8^{hi} DEC205$^{+/+}$ GC B cells (Fig. 1b, c and Extended Data Fig. 1). Decreasing the dose of antigen delivered, by mixing αDEC–OVA with a chimaeric DEC205 antibody carrying the control irrelevant antigen *Plasmodium falciparum* circumsporozoite protein (αDEC–CS), resulted in decreased expansion of

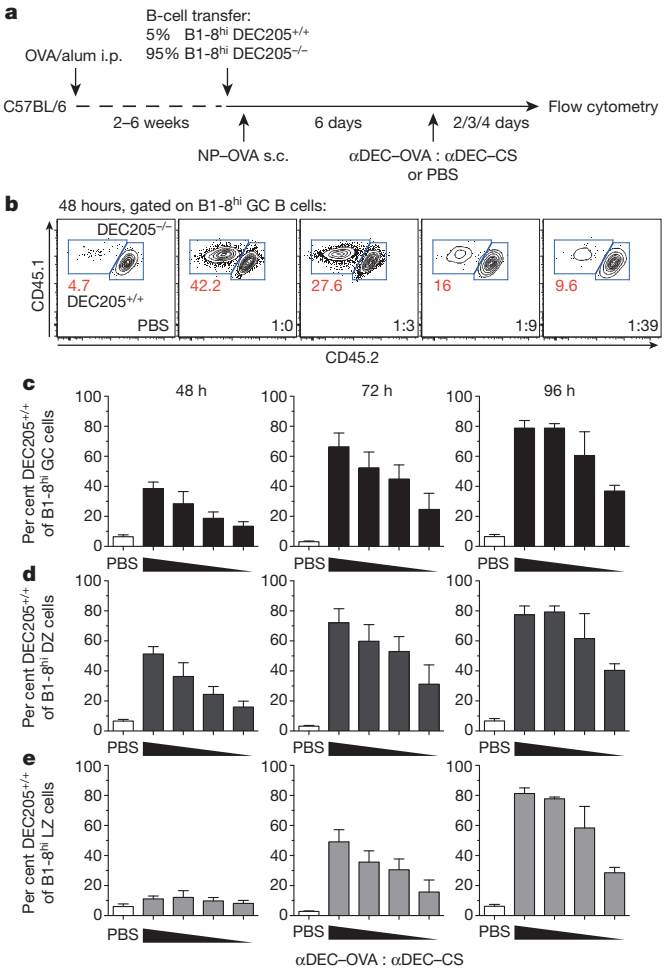


Figure 1 | The amount of antigen captured and presented by GC B cells regulates their expansion. **a**, Protocol for **b**–**e**. i.p., intraperitoneally; s.c., subcutaneously. (1.5–5) × 10^6 B1-8^{hi} DEC205$^{+/+}$ and B1-8^{hi} DEC205$^{-/-}$ B cells (≈ (1.5–5) × 10^5 Igλ$^+$, NP-specific B cells) at a 5:95 ratio were transferred into OVA-primed wild-type mice, which were boosted with NP–OVA. After 6 days, mice were injected with PBS or αDEC–OVA mixed with αDEC–CS at ratios of 1:0, 1:3, 1:9, or 1:39. Lymph nodes were analysed 2, 3 and 4 days after injection. **b**, Proportion of B1-8^{hi} DEC205$^{+/+}$ and B1-8^{hi} DEC205$^{-/-}$ GC B cells 48 h after treatment. **c**–**e**, Mean fraction of DEC205$^{+/+}$ B cells among B1-8^{hi} GC (**c**), DZ (CD86$^-$CXCR4$^+$, **d**), and LZ (CD86$^+$CXCR4$^-$, **e**) cells. Error bars, s.e.m. Data represent 2–3 independent experiments at each time point with a total of 4–6 mice per condition for all time points.

[1]Laboratory of Molecular Immunology, The Rockefeller University, New York, New York 10065, USA. [2]Howard Hughes Medical Institute, The Rockefeller University, New York, New York 10065, USA.

B1-8^{hi} DEC205$^{+/+}$ GC B cells that was proportional to the dose of αDEC–OVA (Fig. 1b, c). Consistent with the idea that pMHCII-mediated selection occurs in the LZ and cell division in the DZ[4,10], selective dose-dependent expansion of B1-8^{hi} DEC205$^{+/+}$ GC B cells was already evident at 48 h in the DZ but only later in the LZ (Fig. 1d, e). In contrast, the B1-8^{hi} DEC205$^{-/-}$ GC B-cell population contracted in proportion to the amount of antigen delivered to the B1-8^{hi} DEC205$^{+/+}$ GC B-cell population (Fig. 1b and Extended Data Fig. 1c). Thus, increasing the amount of cognate antigen presented by a subset of GC B cells to T_{FH} cells leads to their proportional and selective expansion at the expense of GC B cells that present less antigen.

To examine the mechanism by which increased T-cell help leads to selective GC B-cell expansion, we sought to measure cell division in the GC. Traditional dye-based methods to monitor cell division are unsuitable in this context because B cells divide extensively and lose most of the dye before entering the GC. To circumvent this problem, we combined transgenes encoding the tetracycline transactivator (tTA) protein expressed under the Vav promoter and a histone H2B–mCherry fusion protein driven by a doxycycline (DOX)-regulated promoter (tTA–H2B–mCh, Extended Data Fig. 2a)[21,22]. Under steady state conditions, tTA is expressed in haematopoietic cells and induces high levels of H2B–mCh expression (Extended Data Fig. 2b). Administration of DOX represses further H2B–mCh synthesis, and as a result the H2B–mCh dilutes in proportion to cell division (Extended Data Fig. 2c)[23].

To determine whether tTA–H2B–mCh can be used to track antigen-specific GC B-cell division *in vivo*, we repeated the prime-boost protocol described above using B1-8^{hi} tTA–H2B–mCh B cells. As expected, non-proliferative follicular B cells did not dilute H2B–mCh after DOX treatment (Fig. 2a). In contrast, after 36 h, a spectrum of discrete peaks of H2B–mCh expression corresponding to cell divisions became evident among GC B cells, and by 84 h, GC B cells had completely diluted H2B–mCh (Fig. 2a). Thus, tTA–H2B–mCh can be used to monitor cell division in the GC.

To determine whether the amount of antigen captured by GC B cells influences their degree of proliferation, we delivered additional antigen to GC B cells using αDEC–OVA. After 60 h on DOX, B1-8^{hi} DEC205$^{+/+}$ tTA–H2B–mCh GC B cells from mice treated with control αDEC–CS or PBS were nearly evenly distributed among mCh^{Med}, mCh^{Lo} and mCh^{-} groups, representing cells that underwent progressively more cell division. In contrast, after 60 h on DOX, ~50% of GC B cells targeted with αDEC–OVA became mCh^{-} and ~40% were mCh^{Lo} (Fig. 2b, c). Therefore, increased antigen capture and presentation leads to increased rates of cell division by GC B cells.

Concomitant with the increased rate of cell division, there was a change in the zonal distribution of GC B cells. Whereas control cells equilibrated between DZ and LZ at a 2:1 ratio on average, αDEC–OVA-targeted GC B cells were found almost exclusively in the DZ (~90%, Fig. 2d).

To examine cell cycle distribution during selection, we labelled GC B cells with an intravenous pulse of the nucleoside analogs 5-ethynyl-2′-deoxyuridine (EdU) followed 1 h later by 5-bromo-2′-deoxyuridine (BrdU), allowing separation of GC B cells into the earliest (EdU^{-}BrdU^{+}), mid/late (EdU^{+}BrdU^{+}) and post- (EdU^{+}BrdU^{-}) S phase periods of the cell cycle (Fig. 3a and Extended Data Fig. 3). Control GC B cells in early and mid/late-S phases were found in both zones; post-S phase cells were primarily in the DZ (Fig. 3b). This supports the idea that GC B cells can initiate S phase in the LZ and then migrate to the DZ to divide[10]. Moreover, the fact that early S phase cells are distributed at a 2:1 DZ:LZ ratio suggests that, once in the DZ, GC B cells can initiate two cell divisions on average. In contrast, inducing selection by injecting αDEC–OVA shifted early S phase cells among selected GC B cells to the DZ (~86%), whereas slightly fewer early S phase cells were in the DZ (59%) among non-selected cells (Fig. 3b). We conclude that increasing the amount of antigen captured and presented to T_{FH} cells increases the proportion of cells initiating S phase in the DZ. These findings indicate that GC B cells presenting the highest levels of antigen initiate additional cell divisions in the DZ before returning to the LZ.

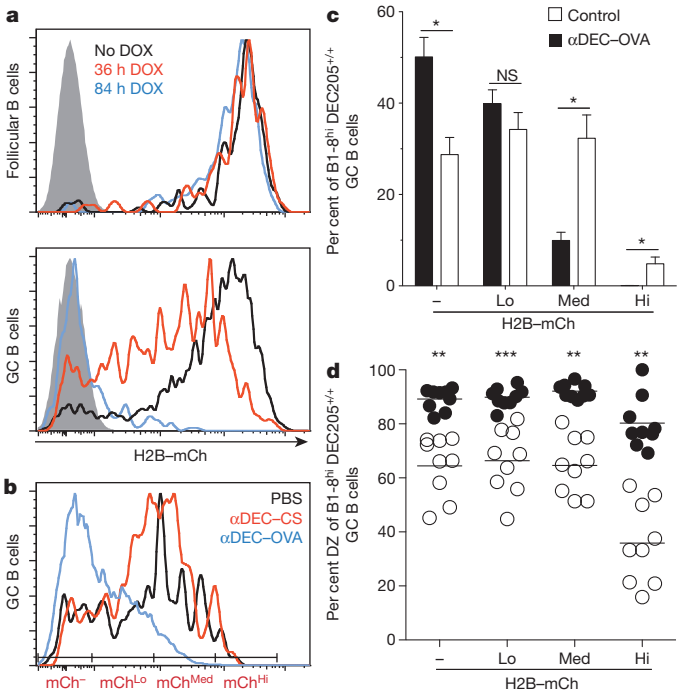


Figure 2 | T-cell help regulates the number of GC B-cell divisions. **a**, H2B–mCh fluorescence among B1-8^{hi} tTA–H2B–mCh B cells within the follicular (upper) or GC (lower) compartments of untreated mice (black) or after 36 (red) or 84 (blue) hours on DOX. Solid grey represents non-fluorescent cells. **b**, H2B–mCh fluorescence among B1-8^{hi} tTA–H2B–mCh GC B cells in mice that received a 5:95 mixture of B1-8^{hi} tTA–H2B–mCh and B1-8^{hi} DEC205$^{-/-}$ B cells and were treated with PBS, αDEC–CS or αDEC–OVA for 72 h and administered DOX for 60 h before analysis. **c**, Mean fraction of B1-8^{hi} tTA–H2B–mCh GC B cells treated as in **b**. Error bars, s.e.m. **d**, Per cent DZ cells among B1-8^{hi} tTA–H2B–mCh GC B cells after treatment with PBS or αDEC–CS (control) or αDEC–OVA for 48 h and DOX for 36 h. Each symbol represents one mouse and lines represent mean values. Data represent 3–4 independent experiments for all time points with 2–3 mice for each condition per experiment. *$P < 0.005$; **$P < 0.001$; ***$P < 0.0001$, two-tailed Mann–Whitney test.

We used photoactivation to test whether selected B cells presenting higher levels of antigen reside longer in the DZ[10]. Primed mice received a mixture of B1-8^{hi} DEC205$^{+/+}$ B cells expressing a photoactivatable green fluorescent protein (B1-8^{hi} PAGFP^{+}) and B1-8^{hi} DEC205$^{-/-}$ B cells. On day 6 of the GC response, we induced selection by injecting αDEC–OVA. DZ cells were photoactivated and lymph nodes were processed for flow cytometry either 0 or 6 h after photoactivation (Fig. 3c and Extended Data Fig. 4). Under control conditions, 57% of photoactivated GC B cells migrated to the LZ after 6 h[10]. In contrast, B1-8^{hi} PAGFP^{+} cells undergoing selective expansion migrated from the DZ to the LZ at half this rate (Fig. 3d, e). Thus, GC B cells undergoing selective expansion as a result of increased antigen presentation reside longer in the DZ, allowing them to undergo additional cell divisions before returning to the LZ. We conclude that T_{FH} cells regulate the number of cell cycles a GC B cell initiates during each passage through the DZ, and that they do so in direct proportion to the amount of antigen captured and presented.

To determine whether variable numbers of cell divisions in the DZ are associated with selection during a polyclonal immune response, we immunized non-immunoglobulin transgenic, tTA–H2B–mCh mice with NP–OVA. Mice were given DOX at day 12.5 after immunization, and 36 h later mCh^{Hi} and mCh^{Lo} GC B cells, representing lower and higher rates of proliferation, respectively, were isolated by cell sorting (Fig. 4a, b). *V_H186.2* family genes were then analysed for the characteristic high affinity anti-NP W33L mutation[24]. We found that 35.9%, of the highly divided GC B cells were W33L^{+}; in contrast, only 6% of the cells that underwent

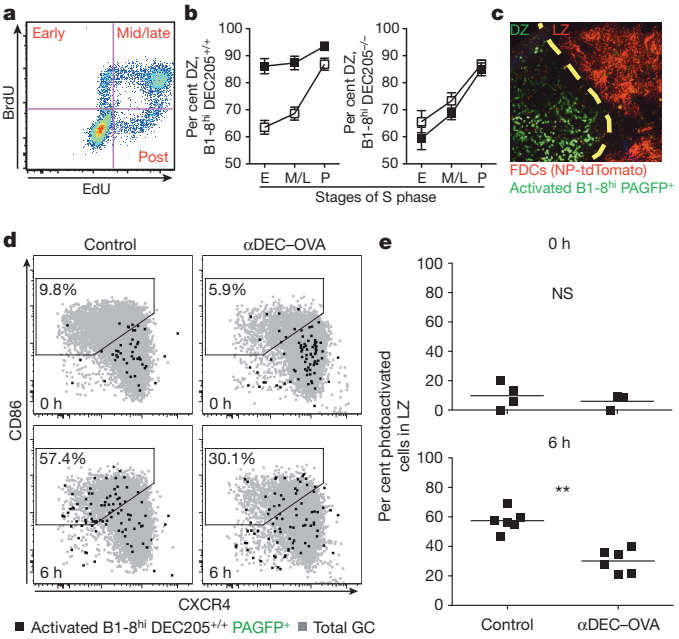


Figure 3 | Selective expansion involves increased S phase initiation in the DZ and longer DZ residence time. **a**, EdU and BrdU incorporation among GC B cells. OVA-primed mice received 5×10^6 B1-8^{hi} $DEC205^{+/+}$ and B1-8^{hi} $DEC205^{-/-}$ B cells at a 15:85 or 50:50 ratio. On day 6 after boosting with NP–OVA, mice receiving the 15:85 or the 50:50 B-cell transfer were injected with αDEC–OVA or PBS, respectively. Two days later, mice received EdU and 1 h later BrdU and were analysed after 30 min (detailed in Extended Data Fig. 3a). **b**, Per cent DZ cells among early (E), mid/late (M/L), and post- (P) S phase cells. Black squares denote αDEC–OVA treatment; white squares denote PBS control. Data represent two independent experiments with 6 or 7 mice per condition in total. Squares indicate mean values; error bars, s.e.m. **c**, **d**, OVA-primed mice received 5×10^6 B1-8^{hi} $PAGFP^+$ and B1-8^{hi} $DEC205^{-/-}$ B cells at a 15:85 ratio or 5×10^6 B1-8^{hi} $PAGFP^+$ B cells alone and were boosted with NP–OVA. For mice receiving mixed B-cell transfers, αDEC–OVA was injected on day 6. On day 8, DZ cells were photoactivated (**c**) and GC B cells were analysed by flow cytometry (detailed in Extended Data Fig. 4a) (**d**). **e**, Per cent LZ cells among photoactivated B1-8^{hi} $PAGFP^+$ GC B cells at 0 or 6 h after DZ photoactivation. Each symbol represents one mouse. Data represent multiple independent experiments. $^{**}P = 0.0022$, two-tailed Mann–Whitney test.

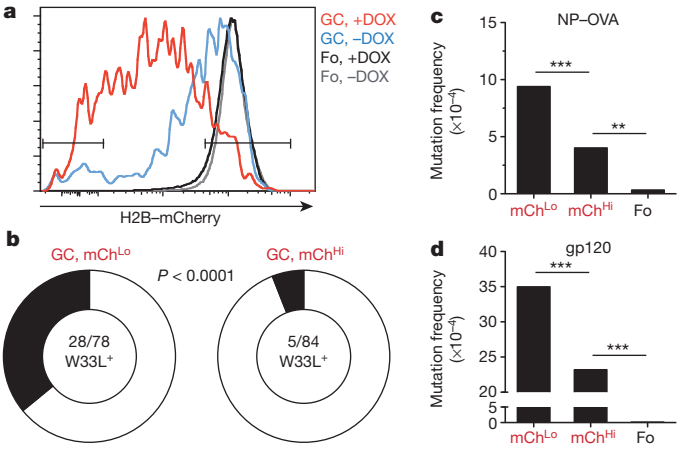


Figure 4 | Increased cell division in polyclonal GCs is associated with higher immunoglobulin affinity and somatic hypermutation. **a**–**d**, tTA–H2B–mCh mice were immunized with NP–OVA (**a**–**c**) or YU2-gp120 (**d**) and administered DOX for 36 h before purification. **a**, Representative histogram displaying mCh^{Hi} and mCh^{Lo} gates. **b**, Frequency of W33L mutation among $V_H186.2$ clones in mCh^{Hi} and mCh^{Lo} GC B cells on day 14 after immunization. The number of $W33L^+$ clones among $V_H186.2$ sequences is shown in the centre of the pie chart and was compared using Fisher's exact test. Data are pooled from 2 independent experiments with 2 mice each. **c**, **d**, Somatic hypermutation in the *J_H4* intron among mCh^{Hi} and mCh^{Lo} GC B cells on day 7 after NP–OVA (**c**) and day 9 after immunization with HIV-1_{YU2} gp120 (**d**). Data in **c** and **d** are pooled from 2 independent experiments with 2–3 mice each. Over 160 clones in **c** and 140 clones in **d** were analysed for mCh^{Hi} and mCh^{Lo} GC B cells and 51 clones in **c** and 94 clones in **d** were analysed for follicular B cells. Fo, follicular B cells. $^{**}P = 0.0033$; $^{***}P < 0.0001$, χ^2 test with Yates correction.

the fewest divisions were $W33L^+$ ($P < 0.0001$, Fig. 4b). Thus, selection of high-affinity GC B cells in a polyclonal response is associated with increased rates of proliferation. Moreover, because each division can introduce mutations that increase or decrease affinity, even high-affinity clones will rapidly partition into subclones that undergo different rates of proliferation, death or differentiation. This is probably reflected in the observation that mCh^{Lo} cells are not all $W33L^+$, and that $W33L^+$ cells are not exclusively in the mCh^{Lo} fraction.

Each round of cell division in the GC is predicted to produce 1 somatic mutation per 10^3 base pairs in immunoglobulin genes[8]. The finding that T_{FH} cells control the number of B-cell divisions per GC cycle indicates that they may also regulate hypermutation. To determine whether there is a correlation between selection, proliferation and somatic hypermutation in the GC, we immunized tTA–H2B–mCh mice with NP–OVA, purified mCh^{Hi} and mCh^{Lo} GC B cells by cell sorting and analysed the intron downstream of *J_H4* for somatic mutations, because this region is targeted for somatic hypermutation but is not subject to selection[25]. GC B cells that had diluted H2B–mCh by undergoing a greater number of divisions were significantly more mutated than those that had undergone fewer divisions ($P < 0.0001$, Fig. 4c).

High levels of mutation are required to produce broadly neutralizing antibodies to human immunodeficiency virus (HIV)-1 and are believed to be an impediment to vaccine development[26]. To determine whether mutations accumulate differentially among GC B cells responding to HIV-1 gp120, we immunized tTA–H2B–mCh mice with HIV-1_{YU2} gp120 and compared GC B cells that had undergone different levels of division. Similar to NP–OVA, the more divided mCh^{Lo} B cells in the anti-gp120 GCs were significantly more mutated than either mCh^{Hi} GC or follicular B cells (Fig. 4d). We conclude that B cells undergo a variable number of divisions in the DZ before returning to the LZ, and that increased cell division is associated with higher immunoglobulin affinity and increased somatic hypermutation.

The GC is a site of intense immunoglobulin diversification and selection, from which high-affinity B cells emerge that seed the memory and plasma cell compartments[3,4,27]. Affinity-based immunoglobulin selection is accomplished iteratively. B-cell clones expanding and mutating in the DZ travel to the LZ to compete for help from T_{FH} cells, with only a fraction of LZ cells selected to return to the DZ to continue cycling[10]. Our experiments indicate that the number of GC B-cell divisions per DZ cycle is variable, ranging from 1 to 6, and is regulated by the amount of antigen captured in the LZ. By capturing more antigen and dividing a greater number of times in the DZ during each interzonal cycle, high-affinity GC B cells outcompete lower affinity cells that capture less antigen and divide fewer times. The magnitude of T-cell help provided in the LZ therefore regulates the behaviour of a DZ cell—inducing selected cells to re-enter the cell cycle a variable number of times before migrating back to the LZ. This finding is consistent with and may explain the observation that the switch from DZ to LZ phenotype is independent of cues received in the DZ[14]. Moreover, the least proliferative GC B cells have the fewest affinity-enhancing mutations and are disproportionately found in the LZ (Figs 2d and 4b). The precise contributions of death and differentiation to the eventual disappearance of these cells from the GC remain to be determined.

An important feature of differential cell division in the dark zone is that each round of division is associated with increased accumulation of somatic hypermutation. Consistent with this idea, highly proliferative germinal centre B cells have both higher immunoglobulin affinity and

a greater number of somatic mutations. Our experiments reveal a feed-forward loop in the germinal centre, whereby somatic hypermutation is greatest among germinal centre B cells whose antibodies provide the highest affinity template for further diversification. This finding may be particularly relevant for designing immunization strategies that elicit broadly neutralizing antibodies to pathogens, like HIV-1, that require exceptionally high levels of mutation.

METHODS SUMMARY

Vav-tTA transgenic mice were bred with Col1A1-tetO-H2B-mCherry mice to produce tTA–H2B–mCh mice[21,22]. B1-8^{hi}, DEC205$^{-/-}$ and PAGFP transgenic mice were described previously[10,19,28]. Resting B cells from indicated mice were purified by magnetic cells sorting (MACS) and transferred intravenously into recipient mice that were primed 2–6 weeks previously by intraperitoneal immunization with 50 μg of OVA in alum. Adoptive transfer recipients were boosted subcutaneously with 25 μg of NP_{14}–OVA. αDEC–OVA and αDEC–CS chimaeric antibodies were produced by transient transfection of 293T cells and were injected subcutaneously as indicated. Polyclonal GCs were generated by immunizing tTA–H2B–mCh mice intraperitoneally and subcutaneously with 50 μg and 12.5 μg of NP_{14}–OVA or 12.5 μg and 6.25 μg of HIV-1_{YU2} gp120 protein in alum, respectively. DOX was administered by injecting 1.6 mg intraperitoneally and 0.2 mg subcutaneously and by supplementing the drinking water with DOX (2 mg ml^{-1}) and sucrose (10 mg ml^{-1}). Intravital imaging and photoactivation were performed as described previously[10,18,29]. Immunoglobulin sequence analysis was performed using MacVector 12.7 and IMGT/V-QUEST.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 January; accepted 1 April 2014.
Published online 4 May 2014.

1. Berek, C., Berger, A. & Apel, M. Maturation of the immune response in germinal centers. *Cell* **67,** 1121–1129 (1991).
2. Jacob, J., Kelsoe, G., Rajewsky, K. & Weiss, U. Intraclonal generation of antibody mutants in germinal centres. *Nature* **354,** 389–392 (1991).
3. Rajewsky, K. Clonal selection and learning in the antibody system. *Nature* **381,** 751–758 (1996).
4. Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annu. Rev. Immunol.* **30,** 429–457 (2012).
5. Allen, C. D., Okada, T., Tang, H. L. & Cyster, J. G. Imaging of germinal center selection events during affinity maturation. *Science* **315,** 528–531 (2007).
6. Hauser, A. E. *et al.* Definition of germinal-center B cell migration *in vivo* reveals predominant intrazonal circulation patterns. *Immunity* **26,** 655–667 (2007).
7. Kocks, C. & Rajewsky, K. Stepwise intraclonal maturation of antibody affinity through somatic hypermutation. *Proc. Natl Acad. Sci. USA* **85,** 8206–8210 (1988).
8. McKean, D. *et al.* Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proc. Natl Acad. Sci. USA* **81,** 3180–3184 (1984).
9. Schwickert, T. A. *et al. In vivo* imaging of germinal centres reveals a dynamic open structure. *Nature* **446,** 83–87 (2007).
10. Victora, G. D. *et al.* Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell* **143,** 592–605 (2010).
11. Allen, C. D., Okada, T. & Cyster, J. G. Germinal-center organization and cellular dynamics. *Immunity* **27,** 190–202 (2007).
12. Oprea, M. & Perelson, A. S. Somatic mutation leads to efficient affinity maturation when centrocytes recycle back to centroblasts. *J. Immunol.* **158,** 5155–5162 (1997).
13. Allen, C. D. *et al.* Germinal center dark and light zone organization is mediated by CXCR4 and CXCR5. *Nature Immunol.* **5,** 943–952 (2004).
14. Bannard, O. *et al.* Germinal center centroblasts transition to a centrocyte phenotype according to a timed program and depend on the dark zone for effective selection. *Immunity* **39,** 912–924 (2013).
15. Dominguez-Sola, D. *et al.* The proto-oncogene *MYC* is required for selection in the germinal center and cyclic reentry. *Nature Immunol.* **13,** 1083–1091 (2012).
16. Jiang, W. *et al.* The receptor DEC-205 expressed by dendritic cells and thymic epithelial cells is involved in antigen processing. *Nature* **375,** 151–155 (1995).
17. Kamphorst, A. O., Guermonprez, P., Dudziak, D. & Nussenzweig, M. C. Route of antigen uptake differentially impacts presentation by dendritic cells and activated monocytes. *J. Immunol.* **185,** 3426–3435 (2010).
18. Shulman, Z. *et al.* T follicular helper cell dynamics in germinal centers. *Science* **341,** 673–677 (2013).
19. Shih, T. A., Roederer, M. & Nussenzweig, M. C. Role of antigen receptor affinity in T cell-independent antibody responses *in vivo*. *Nature Immunol.* **3,** 399–406 (2002).
20. Boscardin, S. B. *et al.* Antigen targeting to dendritic cells elicits long-lived T cell help for antibody responses. *J. Exp. Med.* **203,** 599–606 (2006).
21. Egli, D., Rosains, J., Birkhoff, G. & Eggan, K. Developmental reprogramming after chromosome transfer into mitotic mouse zygotes. *Nature* **447,** 679–685 (2007).
22. Wiesner, S. M., Jones, J. M., Hasz, D. E. & Largaespada, D. A. Repressible transgenic model of *NRAS* oncogene-driven mast cell disease in the mouse. *Blood* **106,** 1054–1062 (2005).
23. Tumbar, T. *et al.* Defining the epithelial stem cell niche in skin. *Science* **303,** 359–363 (2004).
24. Allen, D., Simon, T., Sablitzky, F., Rajewsky, K. & Cumano, A. Antibody engineering for the analysis of affinity maturation of an anti-hapten response. *EMBO J.* **7,** 1995–2001 (1988).
25. Jolly, C. J., Klix, N. & Neuberger, M. S. Rapid methods for the analysis of immunoglobulin gene hypermutation: application to transgenic and gene targeted mice. *Nucleic Acids Res.* **25,** 1913–1919 (1997).
26. West, A. P., Jr *et al.* Structural insights on the role of antibodies in HIV-1 vaccine and therapy. *Cell* **156,** 633–648 (2014).
27. Tarlinton, D. & Good-Jacobson, K. Diversity among memory B cells: origin, consequences, and utility. *Science* **341,** 1205–1211 (2013).
28. Guo, M. *et al.* A monoclonal antibody to the DEC-205 endocytosis receptor on human dendritic cells. *Hum. Immunol.* **61,** 729–738 (2000).
29. Schwickert, T. A. *et al.* A dynamic T cell-limited checkpoint regulates affinity-dependent B cell entry into the germinal center. *J. Exp. Med.* **208,** 1243–1252 (2011).

Acknowledgements We thank S. W. Lowe and D. R. Fooksman for mice; D. Bosque and T. Eisenreich for help with mouse colony management; A. Abadir for protein production; K. Yao for technical help; K. Velinzon for help with cell sorting; D. Mucida and all members of the Nussenzweig laboratory for discussion. Support for the Rockefeller University multiphoton microscope was granted by the Empire State Stem Cell Fund through New York State Department of Health contract C023046. Supported by National Institutes of Health (NIH) Medical Scientist Training Program grant T32GM07739 to the Weill Cornell/Rockefeller/Sloan-Kettering Tri-Institutional MD-PhD Program (A.D.G.); NIH grants AI037526-19 and AI072529-06 (M.C.N.); and the NIH Center for HIV/AIDS Vaccine Immunology and Immunogen Discovery (CHAVI-ID) 1UM1 AI100663-01 (M.C.N). Z.S. is a Human Frontiers of Science Fellow. M.C.N. is an HHMI investigator.

Author Contributions A.D.G. planned and performed experiments and wrote the manuscript. A.D.G. and Z.S. planned and performed photoactivation experiments. M.C.N. planned experiments and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.C.N. (nussen@mail.rockefeller.edu).

LETTER

doi:10.1038/nature13232

Dichloroacetate prevents restenosis in preclinical animal models of vessel injury

Tobias Deuse[1,2,3], Xiaoqin Hua[1,2], Dong Wang[1,2], Lars Maegdefessel[4], Joerg Heeren[5], Ludger Scheja[5], Juan P. Bolaños[6], Aleksandar Rakovic[7], Joshua M. Spin[8], Mandy Stubbendorff[1,2], Fumiaki Ikeno[8], Florian Länger[9], Tanja Zeller[2,10], Leonie Schulte-Uentrop[2,11], Andrea Stoehr[2,12], Ryo Itagaki[1,2], Francois Haddad[8], Thomas Eschenhagen[2,12], Stefan Blankenberg[2,10], Rainer Kiefmann[2,11], Hermann Reichenspurner[2,3], Joachim Velden[13], Christine Klein[7], Alan Yeung[8], Robert C. Robbins[14], Philip S. Tsao[8,15] & Sonja Schrepfer[1,2,3,14]

Despite the introduction of antiproliferative drug-eluting stents, coronary heart disease remains the leading cause of death in the United States[1]. In-stent restenosis and bypass graft failure are characterized by excessive smooth muscle cell (SMC) proliferation[2,3] and concomitant myointima formation with luminal obliteration. Here we show that during the development of myointimal hyperplasia in human arteries, SMCs show hyperpolarization of their mitochondrial membrane potential ($\Delta\Psi m$) and acquire a temporary state with a high proliferative rate and resistance to apoptosis. Pyruvate dehydrogenase kinase isoform 2 (PDK2) was identified as a key regulatory protein, and its activation proved necessary for relevant myointima formation. Pharmacologic PDK2 blockade with dichloroacetate or lentiviral PDK2 knockdown prevented $\Delta\Psi m$ hyperpolarization, facilitated apoptosis and reduced myointima formation in injured human mammary and coronary arteries, rat aortas, rabbit iliac arteries and swine (pig) coronary arteries. In contrast to several commonly used antiproliferative drugs, dichloroacetate did not prevent vessel re-endothelialization. Targeting myointimal $\Delta\Psi m$ and alleviating apoptosis resistance is a novel strategy for the prevention of proliferative vascular diseases.

Balloon injury of Lewis rat aortas triggered an inflammatory response and caused leukocyte infiltration in the SMC-rich media after 48 h, consisting mainly of CD68-positive ($CD68^+$) macrophages and some myeloperoxidase-positive (MPO^+) neutrophils; $CD3^+$ lymphocytes were not observed (Extended Data Fig. 1a–c). Compared to healthy, non-injured aortas, we observed increased phosphorylation of AKT (pAKT), and ERK1, ERK2 (pERK1/2) and $\Delta\Psi m$ hyperpolarization in media cells of injured vessels (Extended Data Fig. 1d, e). A myointima subsequently developed luminal to the internal elastic lamina, which caused progressive luminal obliteration over 28 days (Extended Data Fig. 1f, g). This process was accompanied by leukocyte infiltration and inflammatory cytokine release, which was strong after 7 days and markedly reduced at 28 days (Extended Data Fig. 1h, i).

A humanized model was subsequently developed to study myointima formation longitudinally in human arteries. Balloon-injured human internal mammary arteries (HMAs) were implanted into the abdominal aortic position of T-cell-deficient Rowett nude (RNU) rats (Supplementary Video 1). Myointimal hyperplasia rapidly developed over 4 weeks (Fig. 1a) causing progressive luminal obliteration (Fig. 1c). By histopathology (Fig. 1b) and confocal immunofluorescence (Extended Data Fig. 2a, b), the myointima in the HMA model after 28 days or later closely resembled lesions of diseased human coronary arteries retrieved from autopsy samples. Using human leukocyte antigen class I (HLA I) and rat MHC I antibodies, the human origin of the SMCs within the myointima was confirmed (Extended Data Fig. 2c). Only the mechanical vessel injury was causally related to myointima formation and no relevant xenoantigen-triggered immune activation was observed (Extended Data Fig. 3a, b). Similar to the immunocompetent Lewis rat aortic injury model, we observed accumulation of $CD68^+$ macrophages and MPO^+ neutrophils in HMA vessels after 7 days, which was markedly attenuated by day 28 (Extended Data Fig. 3c). Immune cell infiltration was again accompanied by the elevation of inflammatory cytokines (Extended Data Fig. 3d).

Analysis of cell growth dynamics in HMAs showed a transient but strong increase in proliferative activity within the myointima between 7 and 21 days after injury, accompanied by a persistently low rate of apoptosis (Fig. 1d, e). Proliferation and apoptosis leveled off after 28 days when there was also no further progression of myointimal disease (Fig. 1c, e). Only during the time period of highly positive net proliferation did myointimal cells demonstrate $\Delta\Psi m$ hyperpolarization (Fig. 1f). Within the myointima, cells in the luminal region showed higher proliferative activities and higher $\Delta\Psi m$ than cells closer to the media (Extended Data Fig. 3e).

Platelet-derived growth factor (PDGF) was suspected to be the major driving factor promoting myointimal hyperplasia as it was temporarily increased in injured HMA vessels and PDGF receptor blockade prevented the development of relevant disease (Extended Data Fig. 3f, g). Human vascular SMCs were isolated from fresh HMAs and characterized (Extended Data Fig. 4a, b). PDGF was then shown to induce $\Delta\Psi m$ hyperpolarization in cultured SMCs (Extended Data Fig. 4c), similar to the $\Delta\Psi m$ hyperpolarization previous observed in injured HMAs (Fig. 1f). Thus, mitochondrial $\Delta\Psi m$ hyperpolarization in myointimal SMCs and cultured SMCs coincided with the availability of PDGF. PDGF also caused a phenotype switch in SMCs from a contractile to a dedifferentiated state (Extended Data Fig. 4d).

Mitochondria have been shown to regulate apoptosis via their mitochondrial apoptotic pathway[4]. This involves mitochondrial permeability transition and the release of toxic components such as cytochrome C and caspases[5]. In this context, $\Delta\Psi m$ has an important role in the control of the mitochondrial permeability transition pore, as $\Delta\Psi m$

[1]TSI-laboratory, University Heart Center Hamburg, Martinistraße 52, 20246 Hamburg, Germany. [2]Cardiovascular Research Center Hamburg (CVRC) and DZHK (German Center for Cardiovascular Research), partner site Hamburg/Kiel/Luebeck, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany. [3]Cardiovascular Surgery, University Heart Center Hamburg, Martinistraße 52, 20246 Hamburg, Germany. [4]Department of Medicine, Atherosclerosis Research Unit, Karolinska Institute, CMM L8:03, 17176 Stockholm, Sweden. [5]Department of Biochemistry and Molecular Cell Biology, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany. [6]Institute of Functional Biology and Genomics, University of Salamanca-CSIC, Zacarias Gonzalez 2, 37007 Salamanca, Spain. [7]Institute of Neurogenetics, University of Lübeck, Maria-Goeppert-Straße 1, 23562 Lübeck, Germany. [8]Cardiovascular Medicine and Stanford Cardiovascular Institute, Stanford University, 300 Pasteur Drive, Stanford, California 94305, USA. [9]Institute of Pathology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany. [10]Department of General and Interventional Cardiology, University Heart Center Hamburg, Martinistraße 52, 20246 Hamburg, Germany. [11]Department of Anaesthesiology, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany. [12]Institute of Experimental Pharmacology and Toxicology, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany. [13]Department of Nephropathology, Institute of Pathology, University Hospital Erlangen, Krankenhausstraße 8-10, 91054 Erlangen, Germany. [14]Department of Cardiothoracic Surgery and Stanford Cardiovascular Institute, Stanford University, 300 Pasteur Drive, Stanford, California 94305, USA. [15]Veterans Affairs Palo Alto Health Care System, 3801 Miranda Avenue, Palo Alto, California 94304, USA.

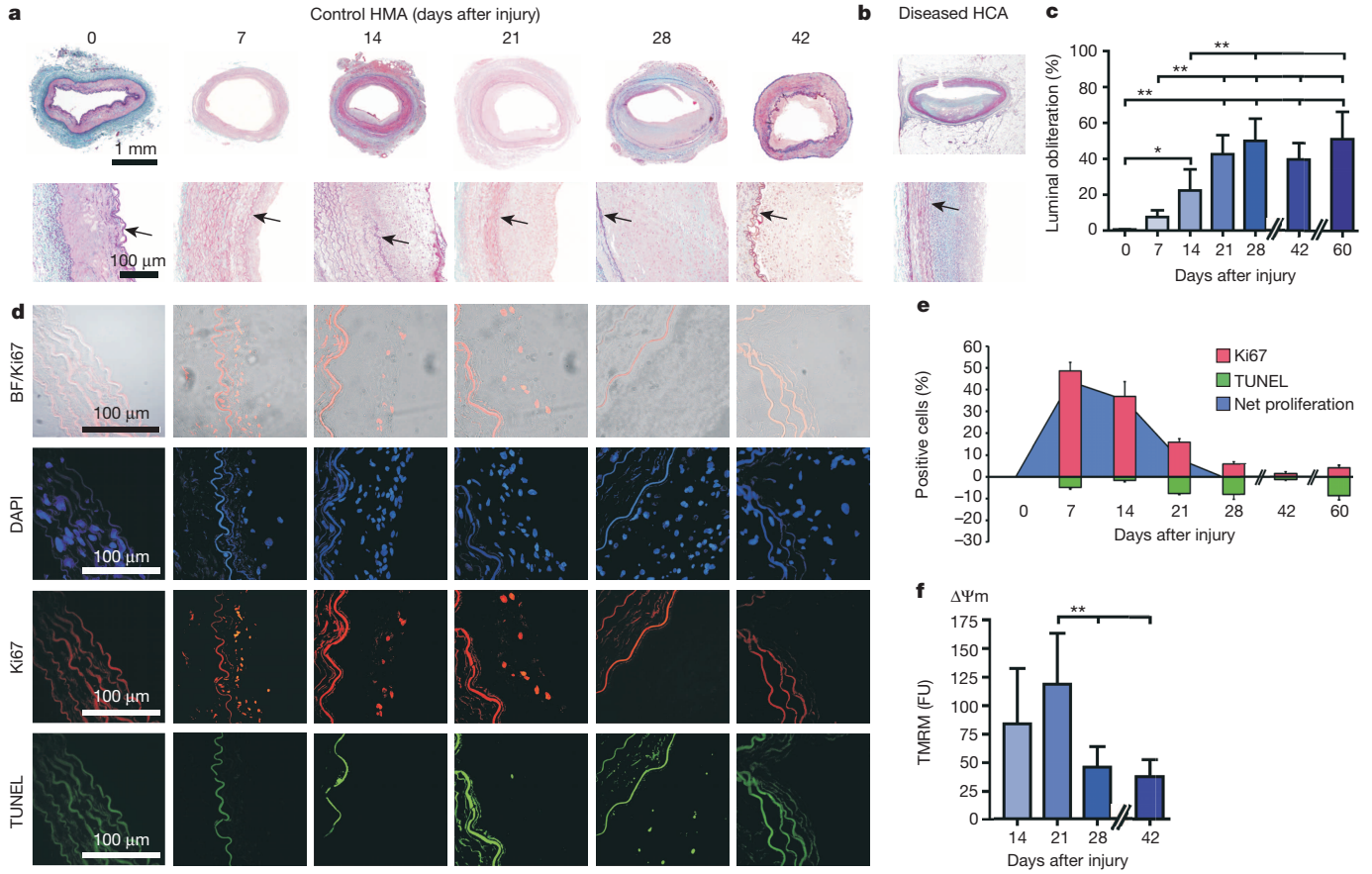


Figure 1 | Chronology and growth dynamics of myointima formation in the HMA model. **a**, Development of myointimal hyperplasia (trichrome; black arrows, internal elastic lamina). **b**, Non-calcified lesions in naturally diseased human coronary artery (HCA). **c**, Luminal obliteration over time (mean ± s.d., $n = 5$ animals (days 0–28), 4 animals (days 42–60), ANOVA with Bonferroni's post-hoc test). **d**, **e**, The timeline of proliferative (Ki67) and apoptotic activity (TUNEL, TdT-mediated dUTP nick end labelling) within the myointima was monitored by immunofluorescence (**d**) and quantified (**e**, mean ± s.e.m., $n = 7$ animals (day 7), 5 animals (day 14), 6 animals (days 21–60)). Net proliferation was calculated as the percentage of Ki67-positive minus TUNEL-positive cells. BF, brightfield. **f**, ΔΨm in HMA myointima (mean ± s.d., $n = 5$ animals (day 14), 9 animals (day 21), 6 animals (days 28–42), ANOVA with Bonferroni's post-hoc test). FU, arbitrary fluorescence units; TMRM, tetramethylrhodamine methyl ester perchlorate assay. $^{*}P < 0.05$, $^{**}P < 0.01$.

hyperpolarization has been suspected to impede pore opening[6,7]. Dichloroacetate (DCA), a rapid-acting small molecule targeting mitochondrial PDKs, has previously been demonstrated to reduce ΔΨm in A549 cells[8]. As elevated ΔΨm and suppressed apoptosis have been observed after injury in our HMA model, we reasoned that DCA would prevent post-injury ΔΨm hyperpolarization, facilitate apoptosis and reduce myointimal growth.

DCA effectively prevented ΔΨm hyperpolarization in PDGF-treated SMCs isolated from either healthy or atherosclerotic vessels (Extended Data Fig. 5a, b). Previously, mitochondrial cytochrome C release and apoptosis induction were shown to be suppressed by hyperpolarized ΔΨm[7,9]. Consistent with this observation, PDGF reduced staurosporine-induced mitochondrial cytochrome C leakage (Extended Data Fig. 5c, d) and rendered SMCs resistant to apoptosis (Extended Data Fig. 5e). DCA both increased cytochrome C release and restored the ability to enter apoptosis (Extended Data Fig. 5c–e).

In vivo, oral DCA did not affect leukocyte infiltration in the media of balloon-injured aortas in immunocompetent Lewis rats, but effectively lowered ΔΨm in the media at 48 h (Extended Data Fig. 6a–d). The accumulation of $CD68^{+}$ macrophages and MPO^{+} neutrophils in the developing myointima was also not affected by DCA (Extended Data Fig. 6e). However, at 7 days, DCA markedly reduced myointimal ΔΨm and permitted apoptosis (Extended Data Fig. 6f, h). At 28 days, ΔΨm in the developed control myointima had already lowered and DCA showed little effect (Extended Data Fig. 6g). Also, proliferation and apoptosis were low at that time point (Extended Data Fig. 6i).

The potential vasculoprotective effect of DCA was then tested in the HMA and human coronary artery (HCA) models (Fig. 2a–d). For the latter, HCAs with minor pre-existing disease, which underwent the same balloon injury and implantation procedure as described for HMA, were used to better reflect the coronary pathophysiology. Oral DCA administration strongly reduced the development of myointima and luminal narrowing in both HMAs (Fig. 2a) and HCAs (Fig. 2d). Similar to untreated HMA controls, the proliferative response to injury in DCA-treated vessels was strongest within the first 14 days and weakened thereafter (Fig. 2b). However, the apoptotic activity was also enhanced and mirrored the proliferative activity at each time point, resulting in a much lower net proliferation. In accordance with the *in vitro* data, we observed significantly lower ΔΨm in the 21-day HMA specimens of the DCA group (Fig. 2c). To exclude the possibility that ischaemia-reperfusion injury or xenogeneic immune interactions might have affected our results, DCA was tested further in the rat aortic (Fig. 2e) and the rabbit iliac artery balloon injury models (Fig. 2f). Again, myointima formation after 28 days was remarkably reduced by DCA in both models. Interestingly, DCA neither demonstrated anti-migratory effects on endothelial cells *in vitro* nor inhibited vessel re-endothelialization *in vivo* (Extended Data Fig. 7 and Supplementary Video 2).

PDKs are the only known targets of DCA and PDKs exclusively phosphorylate and thus inactivate pyruvate dehydrogenase (PDH). It was therefore suspected that PDK knockdown would generate similar biological effects as PDK inhibition by DCA. Although SMCs express PDK1, PDK2 and PDK3 (Extended Data Fig. 8a), PDK2 has the highest affinity

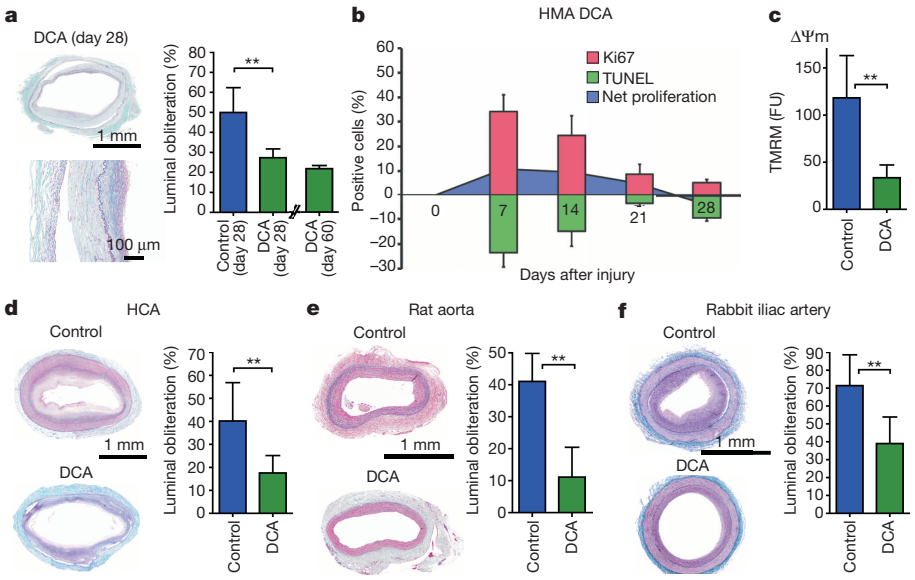


Figure 2 | DCA alleviates myointima formation *in vivo*. **a**, After 28 days, luminal obliteration was significantly less in the DCA group of the HMA model (mean ± s.d., $n = 5$ animals per group, Student's *t*-test). In animals treated for 60 days ($n = 4$), there was no further disease progression. **b**, Myointimal growth dynamics in the HMA DCA group (mean ± s.e.m., $n = 4$ animals (day 7), 5 animals (day 14), 3 animals (day 21), 6 animals (day 28)). **c**, At day 21, ΔΨm was significantly lower in HMA DCA vessels than in control arteries (mean ± s.d., $n = 9$ control animals, $n = 4$ DCA animals, Student's *t*-test). **d**, At day 28, DCA reduced luminal obliteration in the HCA model (mean ± s.d., $n = 8$ control animals, $n = 14$ DCA animals, Student's *t*-test). **e**, **f**, At day 28, DCA effectively alleviated myointima formation in balloon-injured rat aortas (**e**, mean ± s.d., $n = 6$ animals per group, Student's *t*-test) and rabbit iliac arteries (**f**, mean ± s.d., $n = 12$ control animals, $n = 15$ DCA animals, Student's *t*-test). $^{**}P < 0.01$.

to DCA[10] and is likely to mediate the vast majority of the DCA effect. Therefore, PDK2-knockdown SMCs were generated using *PDK2* lentiviral short hairpin RNA (shRNA) (Extended Data Fig. 8b, c) to verify the mechanistic involvement of PDK2 in the DCA effect. Indeed PDK2-knockdown SMCs maintained steady low ΔΨm (Extended Data Fig. 9a) under control conditions and with PDGF stimulation. Also, PDK2-knockdown SMCs maintained increased cytochrome C release and elevated apoptotic rates even during PDGF incubation (Extended Data Fig. 9b–d). In addition, DCA lost its ability to affect ΔΨm in PDK2-knockdown SMCs (Extended Data Fig. 9a). As expected, when PDH (the primary inhibitory target of PDK2) was knocked down (Extended Data Fig. 9e) instead of PDK2, the opposite effect was observed: ΔΨm was permanently elevated, even under control conditions (Extended Data Fig. 9f). DCA again lost its ability to depolarize ΔΨm, indicating that the DCA effect on ΔΨm depended on both PDK2 and PDH.

To assess whether arteries with *ex vivo* PDK2 knockdown would develop similarly reduced myointima as arteries under DCA treatment, PDK2 knockdown was induced in HMAs and HCAs before implantation (Extended Data Fig. 8d–f). Comparable to the DCA group, PDK2-knockdown HMAs showed low net proliferation throughout the 28-day study period (Fig. 3a). ΔΨm hyperpolarization was also effectively prevented and the low ΔΨm of PDK2 knockdown HMAs at day 21 (Fig. 3b) closely resembled the ΔΨm values of DCA-treated HMAs (Fig. 2c). Reduced luminal obliteration was observed in both PDK2-knockdown HMAs (Fig. 3c) and HCAs (Fig. 3d) at day 28.

DCA was ultimately evaluated in a translationally relevant swine model of coronary artery restenosis. Yorkshire swine underwent standardized coronary artery balloon injury under fluoroscopic guidance. After 28 days, DCA-treated animals showed significantly reduced luminal obliteration, proliferation area, and max. proliferation thickness (Fig. 4a–d). In summary, DCA reduced myointima formation in five different preclinical *in vivo* models.

Upon PDGF stimulation of cultured SMCs, we observed an activation of the downstream PI(3)K (phosphatidylinositol-3-OH kinase) and MEK (MAPK kinase) pathways with AKT phosphorylation, ERK1 and ERK2 phosphorylation, hexokinase 2 (HK2) upregulation, and increased HK2–mitochondrial association (Extended Data Fig. 10a–d). HK2 has previously been shown to have a binding site close to VDAC (voltage-dependent anion-selective channel), the most abundant protein of the outer mitochondrial membrane, and HK2 binding was reported to reduce channel conductance[11,12]. Consistent with this previous work, we show that HK2–mitochondrial association coincides with ΔΨm hyperpolarization and apoptosis resistance in our study. Furthermore, displacement of HK2 from its mitochondrial VDAC binding site both reduced ΔΨm and restored the susceptibility to apoptosis (Extended Data Fig. 10e). VDAC closure using the inhibitor DIDS also increased SMC ΔΨm and reduced

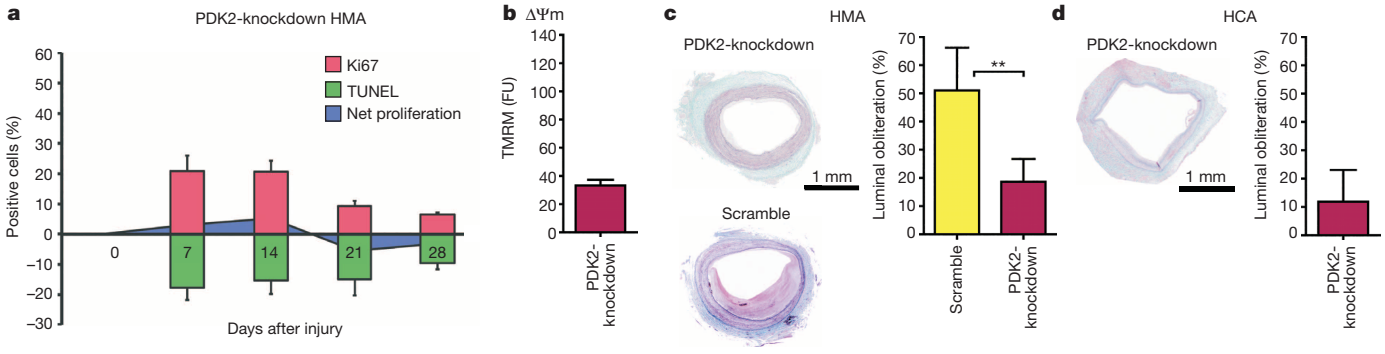


Figure 3 | PDK2 knockdown alleviates myointima formation *in vivo*. **a**, PDK2-knockdown HMAs showed elevated apoptosis rates during the early time points of increased proliferative activity resulting in low myointimal net proliferation (mean ± s.e.m., $n = 3$ (day 7), 4 (day 14), 3 (day 21), 8 (day 28) animals). **b**, At day 21, ΔΨm in PDK2-knockdown HMA vessels was approximately as low as in DCA HMA vessels (mean ± s.d., $n = 4$ animals; see Fig. 2c). **c**, At day 28, luminal obliteration was significantly reduced in PDK2-knockdown HMA compared to scrambled shRNA controls (mean ± s.d., $n = 4$ animals (scramble) and $n = 9$ animals (PDK2-knockdown, Student's *t*-test). **d**, Luminal obliteration was low in PDK2-knockdown HCA vessels at day 28 (mean ± s.d., $n = 3$ animals). $^{**}P < 0.01$.

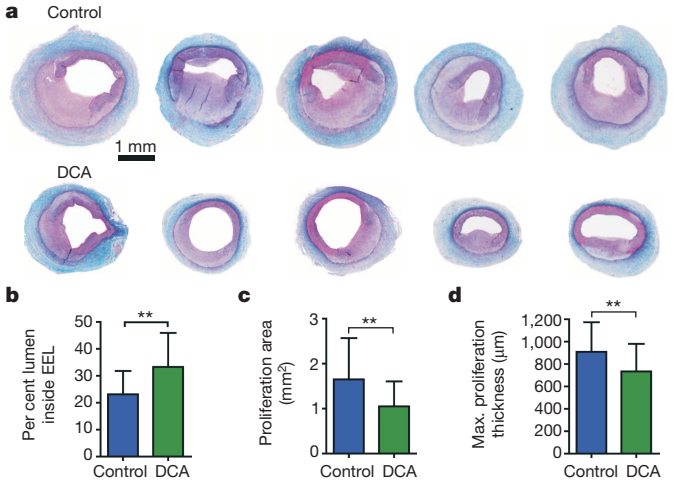


Figure 4 | DCA effectively reduces balloon-injury-induced myointima formation in swine coronary arteries *in vivo*. **a**, After 28 days, the coronary arteries of untreated and DCA-treated Yorkshire swine were retrieved (trichrome). One representative vessel per animal is presented. **b**, The patent lumen inside the external elastic lamina (EEL) was significantly larger in the DCA group (mean ± s.d., $n = 45$ sections per group (three sections per vessel, three coronary arteries per pig, five pigs per group), Student's t-test). Thus, DCA significantly alleviated myointima formation. **c**, **d**, In the region of a ruptured internal elastic lamina, the proliferation area (**c**) and the maximal proliferation thickness (**d**) were significantly smaller with DCA (mean ± s.d., $n = 42$ control sections, $n = 31$ DCA sections, Student's t-test). $^{**}P < 0.01$.

apoptosis (Extended Data Fig. 10f), further supporting the direct mechanistic involvement of VDAC in the regulation of both $\Delta\Psi$m and apoptosis.

DCA and PDK2 knockdown diminished HK2 elevation (Extended Data Fig. 10b) and HK2–mitochondrial association (Extended Data Fig. 10c, d), and maintained low $\Delta\Psi$m despite PDGF (Extended Data Fig. 10g). Notably, it was confirmed that the depolarizing effects of DCA and PDK2 knockdown on $\Delta\Psi$m were ultimately mediated through VDAC, because inhibition of VDAC opening by DIDS was found to neutralize these effects (Extended Data Fig. 10g). However, the mechanistic link between PDK2 inhibition and the decrease in HK2-VDAC binding and reversal of $\Delta\Psi$m hyperpolarization remains elusive.

In healthy vessels, proliferation and apoptosis are very low and balanced[13]. Vessel injury disrupts this homeostasis and triggers an inflammatory state[14–16], which induces temporary $\Delta\Psi$m hyperpolarization in SMCs and drives myointima formation. As shown here, PDK2 repression counteracts temporarily acquired apoptosis resistance and may be a well-tolerated strategy for the prevention of restenosis without interfering with re-endothelialization.

METHODS SUMMARY

Myointimal hyperplasia was induced in HMAs and HCAs by balloon injury and the human arteries were implanted into athymic RNU rats (Crl:NIH-*Foxn1*^{rnu}). Lewis rats and New Zealand rabbits underwent balloon injury of their abdominal aorta or their iliac arteries, respectively. Yorkshire swine underwent percutaneous transluminal coronary angioplasty to achieve a 1.3–1.35:1 balloon–artery overstretch. Rats and rabbits were treated with DCA 0.75 g l^{-1} in drinking water; swine received DCA at 25 mg kg^{-1} three times a day. Confluent SMCs were incubated with PDGF (50 ng ml^{-1}) for 48 h and/or DCA (5 mM; starting 2 days before PDGF) if not indicated otherwise. *PDK2* shRNA lentiviral particles (Santa Cruz) were used to knock down PDK2 in SMCs *in vitro* or *ex vivo* in HMAs or HCAs before implantation. PDH was knocked down in SMCs using short interfering RNA (siRNA) (Santa Cruz). For organ chamber experiments, fresh vascular rings were pre-constricted with prostaglandin F2α and nitroglycerin or acetylcholine was added for relaxation studies. To measure $\Delta\Psi$m, freshly isolated tissues or cells were incubated with tetramethylrhodamine methyl ester perchlorate (TMRM, Invitrogen) and Hoechst 33342 (Invitrogen). Fluorescence signals were quantified by confocal microscopy (Perkin Elmer). SMCs were treated with staurosporine (50 nM, Sigma) for 18 h to induce apoptosis. ApopTag apoptosis detection kit (Merck Millipore) was then used to label apoptotic cells.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 October 2013; accepted 6 March 2014.
Published online 20 April 2014.

1. Lloyd-Jones, D. *et al.* Heart disease and stroke statistics—2009 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation* **119,** 480–486 (2009).
2. Dzau, V. J., Braun-Dullaeus, R. C. & Sedding, D. G. Vascular proliferation and atherosclerosis: new perspectives and therapeutic strategies. *Nature Med.* **8,** 1249–1256 (2002).
3. Novak, K. Cardiovascular disease increasing in developing countries. *Nature Med.* **4,** 989–990 (1998).
4. Green, D. R. & Reed, J. C. Mitochondria and apoptosis. *Science* **281,** 1309–1312 (1998).
5. Hengartner, M. O. The biochemistry of apoptosis. *Nature* **407,** 770–776 (2000).
6. Zamzami, N. & Kroemer, G. The mitochondrion in apoptosis: how Pandora's box opens. *Nature Rev. Mol. Cell Biol.* **2,** 67–71 (2001).
7. Halestrap, A. P. What is the mitochondrial permeability transition pore? *J. Mol. Cell. Cardiol.* **46,** 821–831 (2009).
8. Bonnet, S. *et al.* A mitochondria–K$^+$ channel axis is suppressed in cancer and its normalization promotes apoptosis and inhibits cancer growth. *Cancer Cell* **11,** 37–51 (2007).
9. Bernardi, P. Modulation of the mitochondrial cyclosporin A-sensitive permeability transition pore by the proton electrochemical gradient. Evidence that the pore can be opened by membrane depolarization. *J. Biol. Chem.* **267,** 8834–8839 (1992).
10. Roche, T. E. & Hiromasa, Y. Pyruvate dehydrogenase kinase regulatory mechanisms and inhibition in treating diabetes, heart ischemia, and cancer. *Cell. Mol. Life Sci.* **64,** 830–849 (2007).
11. Azoulay-Zohar, H., Israelson, A., Abu-Hamad, S. & Shoshan-Barmatz, V. In self-defence: hexokinase promotes voltage-dependent anion channel closure and prevents mitochondria-mediated apoptotic cell death. *Biochem. J.* **377,** 347–355 (2004).
12. Pastorino, J. G., Shulga, N. & Hoek, J. B. Mitochondrial binding of hexokinase II inhibits Bax-induced cytochrome c release and apoptosis. *J. Biol. Chem.* **277,** 7610–7618 (2002).
13. Clarke, M. C. *et al.* Apoptosis of vascular smooth muscle cells induces features of plaque vulnerability in atherosclerosis. *Nature Med.* **12,** 1075–1080 (2006).
14. Bernal-Mizrachi, C. *et al.* Vascular respiratory uncoupling increases blood pressure and atherosclerosis. *Nature* **435,** 502–506 (2005).
15. Mitra, A. K., Del Core, M. G. & Agrawal, D. K. Cells, cytokines and cellular immunity in the pathogenesis of fibroproliferative vasculopathies. *Can. J. Physiol. Pharmacol.* **83,** 701–715 (2005).
16. Newby, A. C. & Zaltsman, A. B. Molecular mechanisms in intimal hyperplasia. *J. Pathol.* **190,** 300–309 (2000).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Pahrmann for performing all cell cultures and for her technical assistance. We thank J. Thoms for performing immunoblots, H. Wiebold for assistance in organ chamber experiments, J. Lyons and her team for her assistance with the swine study, and S. Ehret, A. Deng and M. Resch for their technical assistance. We thank the UKE Imaging Facility (UMIF, B. Zobiak) and the UKE Animal Facility. Ethicon (Norderstedt, Germany) provided surgical suture materials. We also thank A. Treszl and G. Schoen for their statistical analyses. This study was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft (DFG), SCHR992/3-1 and SCHR992/4-1 to S.S.), the International Society for Heart and Lung Transplantation (ISHLT, to S.S.), the Förderverein des Universitären Herzzentrums Hamburg (to S.S.), the Hermann and Lilly Schilling Foundation (to C.K.), the MINECO (SAF2013-41177-R, to J.P.B.) and the NIH (NIH 1R01HL105299, to P.S.T.).

Author Contributions T.D. designed the studies, performed surgical procedures, analysed the data and wrote the manuscript. X.H. and D.W. performed confocal immunofluorescence, immunohistochemistry, TMRM assays *in vivo* and *in vitro*, endothelial cell assays, the histopathological studies, analysed data and edited the manuscript. L.M. participated in designing experiments, analysed data, edited the manuscript and secured funding. J.H., L.S., J.P.B., A.R., J.M.S. and C.K. participated in designing experiments, analysed data and edited the manuscript. M.S. performed surgical procedures in rats, contributed to molecular biology experiments and flow cytometry. F.I. and A.Y. contributed to the swine coronary artery procedures and edited the manuscript. F.L. contributed to human coronary artery pathohistology data. T.Z. performed organ chamber experiments. L.S.-U. and R.K. contributed to the rabbit experiments. A.S. performed PCR experiments and analysed data. R.I. performed histology and immunological experiments. F.H. edited the manuscript. T.E., S.B. and H.R. edited the manuscript. J.V. performed histologic analyses. R.C.R. contributed to the study design and edited the manuscript. P.S.T. designed experiments, analysed data and edited the manuscript. S.S. designed the studies, performed surgical procedures in rats and rabbits, ran molecular biology experiments *in vitro* and *in vivo*, edited the manuscript and secured the funding.

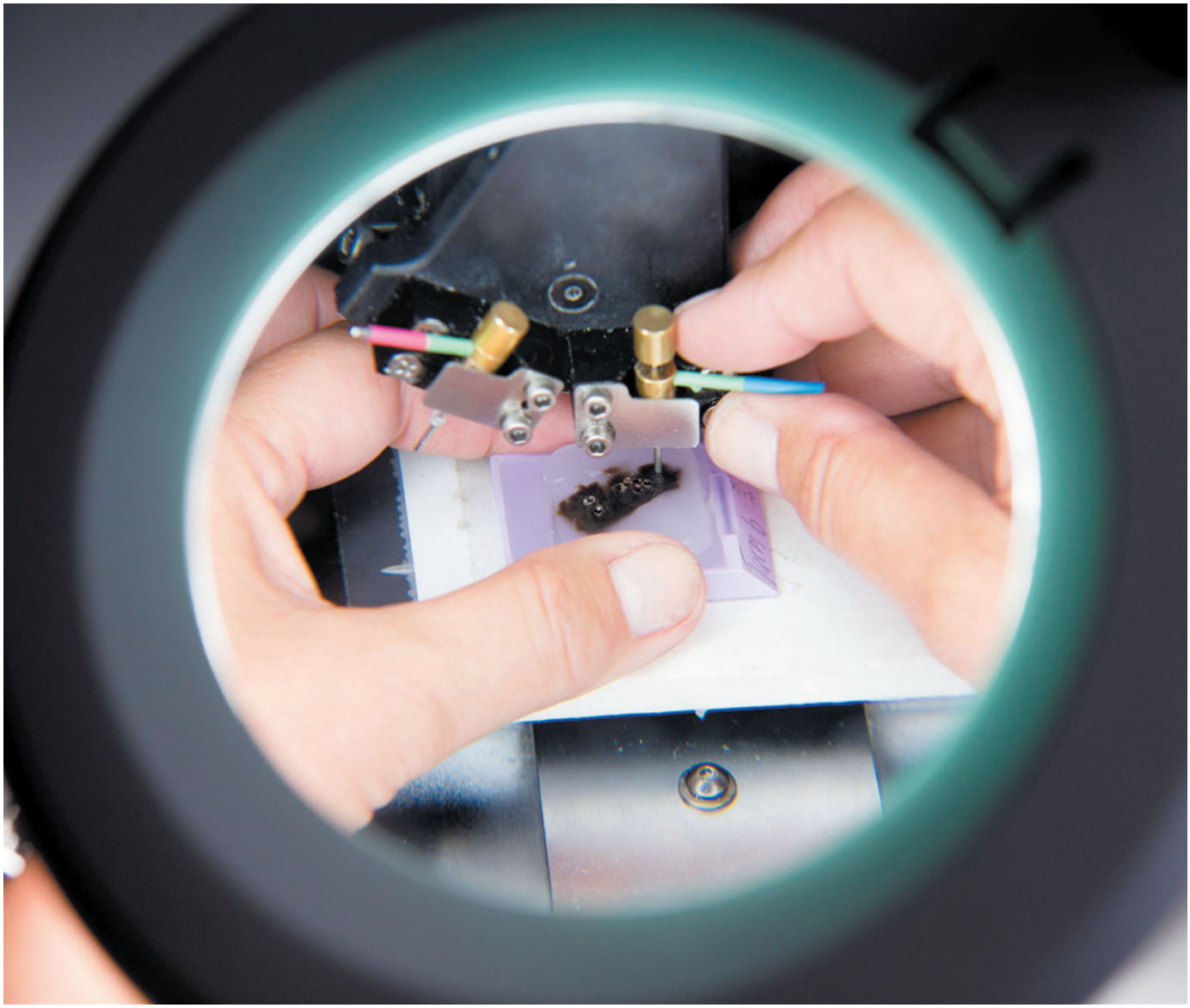
Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.S. (schrepfer@stanford.edu).

TECHNOLOGY FEATURE

AN ATLAS OF EXPRESSION

The first draft of the complete human proteome has been more than a decade in the making. In the process, the effort has also delivered lessons about technology and biology.

HUMAN PROTEIN ATLAS



A researcher removes small sections from a tissue sample to analyse its protein expression for the Human Protein Atlas project.

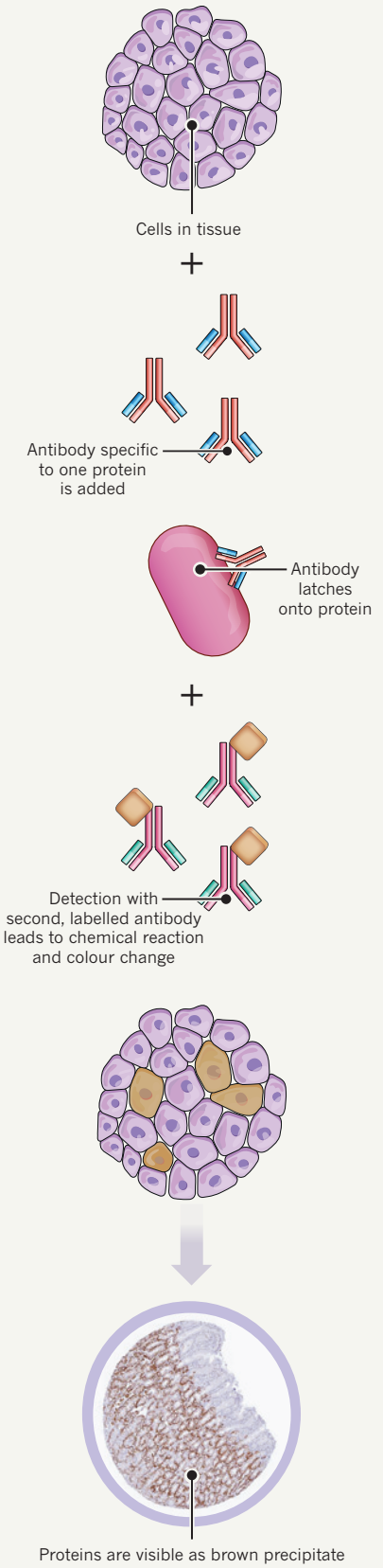
BY VIVIEN MARX

In 2003, two years after the Human Genome Project published a first-draft sequence of the roughly 20,000 genes that define *Homo sapiens*, a group of Swedish researchers launched the Human Protein Atlas (HPA) — a large-scale effort to map where the proteins encoded by those genes are expressed in the body's tissues and cells.

Few proteins had been localized in that way, explains Mathias Uhlén, a protein researcher at the Royal Institute of Technology in Stockholm, and the HPA's principal investigator. A comprehensive atlas of the human 'proteome' would set the stage for more-sophisticated study of protein function, he says. It would reveal the array of membrane proteins, which ferry molecules in and out of cells, and expose the 'secretome' — the proteins secreted by cells in health and in disease. It would guide exploration into the physiological impact of genetic variation. And it would help drug developers to predict where a candidate drug might interact with a protein or cause side effects (see 'Uses for the Human Protein Atlas').

EXPRESS YOURSELF

Cells in a given tissue may look the same but express different proteins. With the help of tissue stains and antibodies, proteins can be tagged and rendered visible.



MAP QUEST

Uses for the Human Protein Atlas

- Look at the spatial distribution of a protein in tissues or cells of interest.
- Study a protein in many different tissues.
- Search for protein-expression patterns that match those of a protein of interest.
- Select cell lines for experiments on the basis of their RNA-expression data and choose corresponding antibodies.
- Begin early studies of protein function.
- Compare normal and cancerous tissues to help to find biomarkers.
- Compare with protein-location data sets generated in other labs.
- Find antibodies for certain applications.
- Compare protein expression in cancerous and normal tissues.
- Use as a teaching tool for cell biology and histology.

The project's annual releases of preliminary maps and data have already yielded surprises, says Fredrik Pontén, a pathologist at Uppsala University in Sweden and a co-founder of the HPA. For example, the maps show that some 3,500 genes encode proteins specific to one tissue type or a small group of tissues, and that the testis contains one-third of those proteins — more than in any other organ[1] (see 'Our proteins, ourselves').

In November, the team plans to upload data completing the first draft of the human proteome. In total, it includes nearly 15 million high-resolution micrographs of stained tissues and cells and presents data on about 80% of human proteins. This information was gleaned from 46 human cell lines and tissue samples from 360 people — 44 normal tissue types from 144 people, and the 20 most common types of cancer, from 216 people. "What's most exciting is the scale of it," says Pontén. Another project — dubbed the subcellular protein atlas — will locate proteins within cells, and will be completed next year.

The samples used for the HPA came from people who agreed to donate tissue while they were being treated for various conditions. In accordance with Swedish research-ethics guidelines, all samples have been anonymized so that they cannot be traced to their donors.

The HPA is not the only protein-mapping venture. Other efforts include the Human Proteome Project, which is loosely affiliated with the HPA but uses some different strategies, and projects in individual labs or among groups of researchers. Two drafts of the human proteome, based on mass spectrometry, are presented in this issue of *Nature*[2,3]. In this approach, tissues are processed and their proteins broken into fragments. The fragments are ionized and then separated according to their mass and charge, which helps to measure, sort and identify the proteins.

The HPA is the only large-scale mapping effort based on antibody-profiling methods, in which chemical stains and antibodies are used to locate and identify proteins in tissues (see 'Express yourself'). The project's scientists used a 'brute force' approach that includes some automation but requires many manual steps. Applying their techniques to the many proteins in the body is daunting — the sum of human proteins exceeds the number of genes by far, and could run in the millions.

With funding from the Knut and Alice Wallenberg Foundation in Stockholm, the HPA researchers — 140 scientists in 13 groups working mainly in Stockholm, Uppsala or Mumbai, India — divided up the work. They scaled up standard proteomics approaches, such as immunohistochemistry, in which antibodies and stains are used to visualize proteins in tissue samples, and western-blot assays to check the specificity of antibodies.

There is a need for such corroboration — often, different approaches yield contradictory results. The researchers have spent time, effort and resources to ensure that the antibodies they use work as expected, and to develop reliable ways to interpret and classify the stained samples, Uhlén says.

SCALING UP

The researchers have also struggled with the general lack of automation in proteomics, says Caroline Kampf, a proteomics researcher and director of the HPA's Uppsala site, who joined the project when it began. That situation has improved over the past decade.

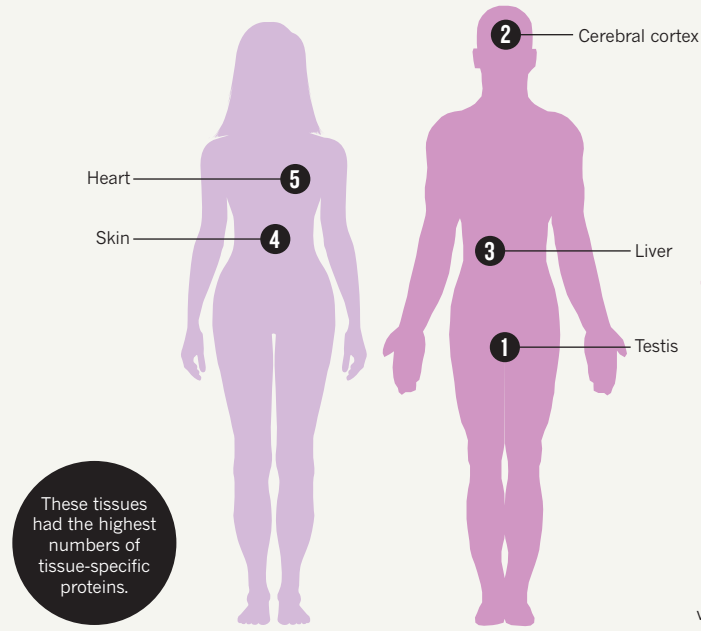
"The Human Protein Atlas has delivered many lessons about tissue staining and large-scale projects."

The task of locating proteins begins with preparation of tissue samples that Kampf and her team receive from Uppsala Biobank, which oversees management of the tissues. The scientists consult research literature and protein databases and also study the tissue's messenger RNA (mRNA), a molecule that carries the information used to manufacture proteins from a DNA sequence.

In 2012, the researchers began using high-throughput mRNA sequencing, or 'RNA-seq', which has enabled them to more quickly obtain data about the set of genes expressed as proteins in a given tissue. This approach has helped to validate results from antibody-based

OUR PROTEINS, OURSELVES

The Human Protein Atlas is based on data from 46 cell lines and from samples of normal and cancerous tissues from 360 people. The five tissues found to have the highest numbers of tissue-specific proteins are shown.



15 MILLION
high-resolution micrographs of stained tissue and cells

Coverage of more than
17,000
protein-encoding genes

37,000
validated antibodies

HUMAN PROTEIN ATLAS

Mathias Uhlén directs the Human Protein Atlas.

tissue profiles, says Kampf.

Tissue preparation involves preserving tissue samples in blocks of paraffin and then processing them into microarrays, groups of tiny tissue samples arranged in a grid to enable scientists to test for the presence of many proteins. Only a fraction of the tissue is used for the microarray; the rest is kept in a repository[4]. For the microarray, cylindrical 'cores' of tissue of around 1 millimetre in diameter are punched out of the paraffin blocks. These cores are embedded in rows and columns in another paraffin block, which is then sliced into thin sections and placed on a slide for staining. The researchers produce around 100 slides a day.

TECH SUPPORT

Although production of tissue microarrays can be automated, says Kampf, plenty can go wrong, and skilled technicians are needed for troubleshooting and for their craftsmanship. For example, she says, a technician needs the right touch to know when to stop tissue punching, or when the punch is stuck. The degree of intervention required depends in part on the sample — differences in texture mean that some tissue types are more challenging than others. Skin, for example, is tougher than fatty breast tissue. To accelerate production, the scientists group similar tissue types for similar processing steps. Some steps still require manual labour. "But when punching cell lines or punching cancers, you can more easily let a machine do it because [the tissue has] a more homogeneous composition," Kampf says.

Once a tissue sample is on the slide, it is treated with reagents that bind specifically to one protein and not another. Another challenging aspect of the procedure is that the concentration of a protein in a given tissue can vary within and between samples. In the body, protein abundance varies by as much as 1-million-fold. The abundant proteins can swamp out the more scarce ones and make them hard to detect.

There are many types of affinity reagents, both biological and synthetic. The HPA uses polyclonal antibodies, which recognize portions of specific proteins. They are produced by injection of an antigen into laboratory animals and later harvesting of the antibodies the animals produce in response. These antibodies are not identical to one another.

"It is always easier to interpret your results if you spent time and effort validating reagents."

The HPA teams have developed methods of antigen design and antibody purification that maximize the specificity with which an antibody latches on to a protein, raising the probability that it will locate one protein only. To date, researchers have validated 37,000 antibodies for the project.

The antibodies developed in the course of the HPA project are available through Atlas Antibodies, an HPA spin-out in Stockholm. The company now sells 17,000 polyclonal antibodies developed by the HPA, and expects to add another 2,000 by November, says Marianne Hansson, the firm's chief executive officer. What is special about this collection of antibodies is that they are produced in a uniform fashion and are 'proteome-wide', says Henrik Wernérus, chief scientific officer at Atlas.

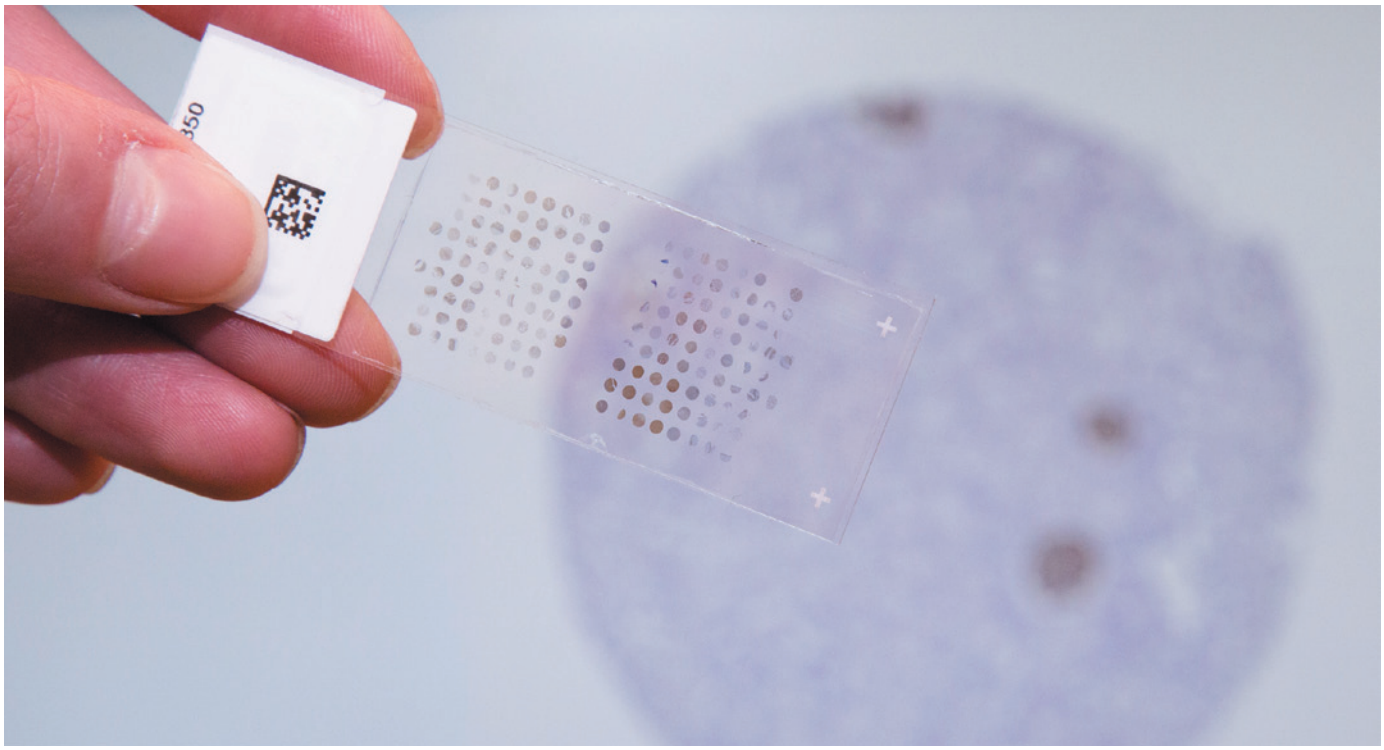
One validation test for antibodies is the tried-and-true western blot, in which proteins are propelled through a gel by an electric field and separated into fragments of different molecular weights. The fragments are then transferred to a material such as nitrocellulose paper and probed with a primary antibody, which binds the protein of interest, followed by a secondary antibody that binds the primary one and bears a fluorescent or enzymatic tag to enable detection of the protein.

Each step in the mapping process takes time, and there are bottlenecks. For example, the project's tissue slides quickly piled up. An automated system cannot readily emulate a pathologist changing the focus on a microscope to study a tissue's protein-expression patterns. Since the HPA began, automated slide readers have become available and are now used to supplement the educated eyes of the project's pathologists.

Kampf says that the first slide scanners her team used could handle around 10 slides a day. But she needed to scan more than 100 slides a day. Now, she says, a scientist can "go home, and when you are back in the morning, they are all there".

But even as automated slide scanning emerged, Kampf and her staff still had to load the slides manually. And the first auto-loaders were far from perfect. Kampf recalls coming into the lab after a coffee break to find the floor covered in glass because an instrument's

HUMAN PROTEIN ATLAS



HUMAN PROTEIN ATLAS

A slide of tissue sections to be stained and incubated with antibodies to see which proteins are expressed.

grippers had smashed the slides against the wall. She spent many an evening on the phone with companies in various time zones to address technical issues.

Kampf says that scaling up is not just about data generation and analysis. "You encounter many other things that you haven't expected," she says. The staining patterns in tumour tissue, for example, can be difficult to interpret, in part because such tissues contain a mix of normal and diseased cells. Cancer cells also show variation — within a tumour as well as between tumour types and patients. Given this variability, data need to be carefully evaluated.

OBJECT LESSONS

Pontén says that the HPA has delivered many lessons about immunohistochemistry and large-scale projects. The scientists have found, for example, that RNA-seq data can help to shore up the results of immunohistochemistry. And they have grappled with differing results between immunohistochemistry and western blots. Whereas the proteins analysed in a western blot have been denatured into two-dimensional fragments, those probed in tissue samples are more likely to retain their three-dimensional (3D) structure. That distinction could lead to differing results, as some antibodies recognize proteins only in their 3D form.

One solution has been to produce the antibodies in a different way, says Pontén. The antigens used conventionally to create antibodies are peptides, short strings of amino acids. In their approach to making antigens, the HPA team designed longer protein fragments. The fragments are 50–150 amino acids in length and are selected to have the highest likelihood of yielding unique proteins, and to contain two or three epitopes — the sites that antibodies recognize on a protein. This multiplicity increases the likelihood that the resulting antibodies will work in western blots, immunohistochemistry and other techniques, Pontén says.

Caroline Kampf says that slide scanners could once handle 10 slides a day but now can handle more than 100.

The results of these assays depend not only on antibody affinity, but also on the relative abundance of a target protein, says Uhlén. Because it can be tough to detect low-abundance proteins, results from antibody staining should be validated with a different antibody for the same protein, for example, or by checking the RNA-seq data to be sure that the RNA that gives rise to the protein in question is present in the tissue, he says.

Antibody validation has "turned out to be even more important than expected", says Emma Lundberg, who directs the subcellular protein atlas project, which shows where in a cell a particular protein is present. Antibodies can have affinity for proteins other than their targets, leading to cross-reactivity that can confuse protein-mapping results. "In the end, it is always easier to interpret your results if you spent some time and effort validating your reagents," she says.

Kampf says that she has, over the years, received many e-mails from scientists asking when their favourite proteins will be published in the atlas. All she could tell them was to be patient. The process, from designing an antigen to having fully validated results ready to upload to the HPA website, takes 9–12 months on average.

The researchers still have to refine the map, says Pontén. There are many low-abundance proteins, among others, to find. And some proteins have been overlooked because they are expressed only at certain times during development, says Pontén. Others have gone undetected because they are present in tissues that the HPA has only begun to collect, such as the retina.

The team hopes that the research community will help with spotting inaccuracies in the map. Pontén says that it will take at least another five years of curation to offer the research community HPA data that have been analysed and assessed to the level of 'textbook' authority. ■

Vivien Marx *is technology editor for* Nature *and* Nature Methods.

1. Fagerberg, L. *et al. Mol. Cell Proteomics* **13,** 397–406 (2014).
2. Kim, M.-S. *et al. Nature* **509,** 575–581 (2014).
3. Wilhelm, M. *et al. Nature* **509,** 582–587 (2014).
4. Kampf, C. Olsson, I., Ryberg, U., Sjöstedt, E. & Pontén, F. *J. Vis. Exp.* **63,** e3620 (2012).

nature OUTLOOK

CANCER

29 May 2014 / Vol 509 / Issue No. 7502

Cover art: Brendan Monroe

For millennia, humans have met their demise through violence, accidents and a fearsome array of infectious diseases. In 1900, the leading causes of death in the United States were pneumonia, influenza and tuberculosis. A century later, they are heart disease and cancer.

Antibiotics and other modern medicines have reduced the lethality of the microbial illnesses that killed our ancestors. Still, we all die of something. So we now find lying in wait for us scores of disorders characterized by the uncontrolled growth of cells. More than forty years since 'war' was declared on cancer, malignancy still casts a shadow over humanity: in 2012, 15% of deaths worldwide were attributable to cancer (page S50). The toll will almost certainly rise in the decades ahead, especially as developing countries adopt Western diets and lifestyles (S64).

This Outlook presents an overview of the current battles against cancer. We examine advances in personalized treatments (S52), nanodevices that will precisely deliver drugs to tumours (S58) and the radical changes that may be needed in clinical research as a result (S55). We explain how the terabytes of data produced by cancer research could be too much of a good thing until we figure out better ways to manage the information (S66 and S68). Clues to potential therapies may lie in an animal that is close to cancer-free (S60), but prevention seems daunting given how much of the environment is potentially carcinogenic (S62). And even as scientists begin to solve the great puzzles concerning cancer, three fundamental mysteries are proving tough to crack (S69).

To deliver this broad view of cancer widely, this Outlook is being published in both *Nature* and *Scientific American* — a collaboration that we expect to be the first of many.

We are pleased to acknowledge the financial support of Celgene Corporation in producing this Outlook. As always, *Nature* has sole responsibility for all editorial content.

Herb Brody
Supplements Editor

CONTENTS

COLLECTION

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at http://www.nature.com/advertising/resources/pdf/outlook_guidelines.pdf

CITING THE OUTLOOK
Cite as a supplement to *Nature*, for example, Nature **Vol XXX**, No. XXXX Suppl, Sxx–Sxx (2014). To cite previously published articles from the collection, please use the original citation, which can be found at the start of each article.

VISIT THE OUTLOOK ONLINE
The *Nature Outlook Cancer* supplement can be found at http://www.nature.com/nature/outlook/cancer

All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES
For UK/Europe (excluding Japan): Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea:Nature Publishing Group — Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

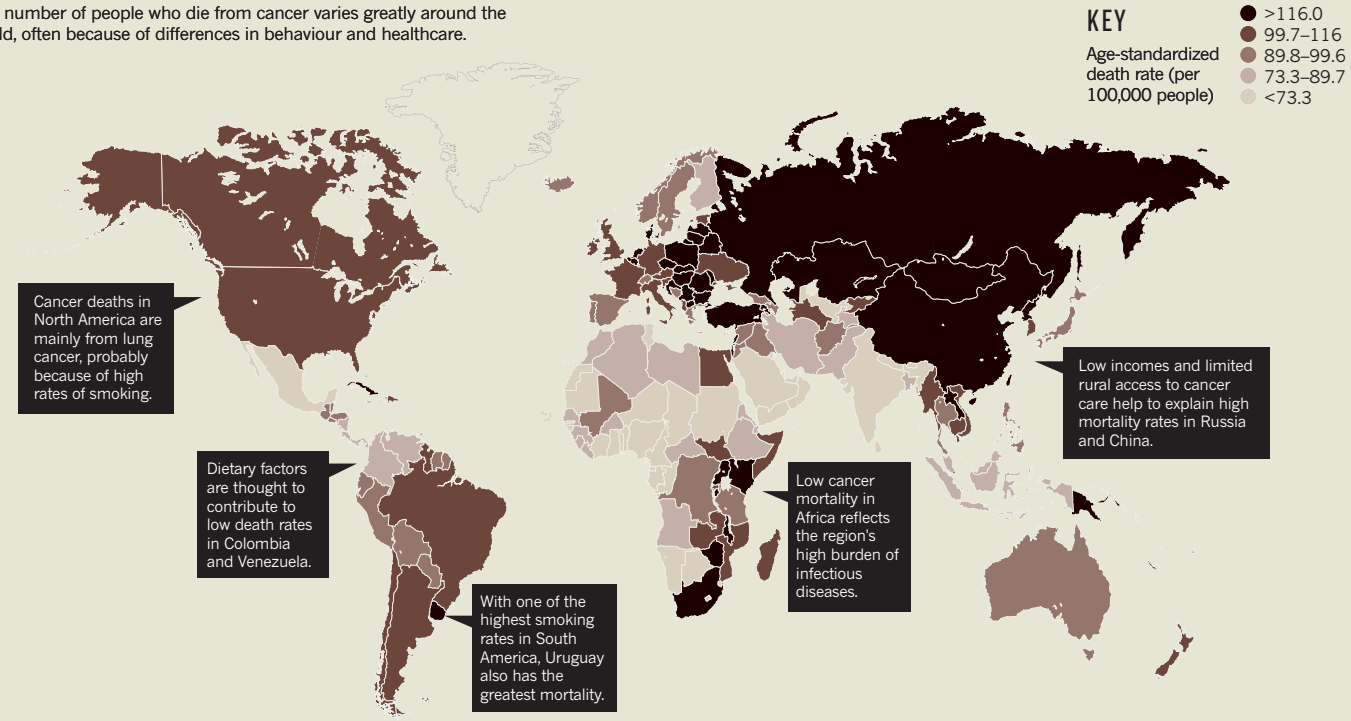
CUSTOMER SERVICES
Feedback@nature.com

ATTACKING AN EPIDEMIC

Despite a huge amount of funding and research, regional and individual differences in cancer trends make it a hard disease to wipe out. By **Mike May**.

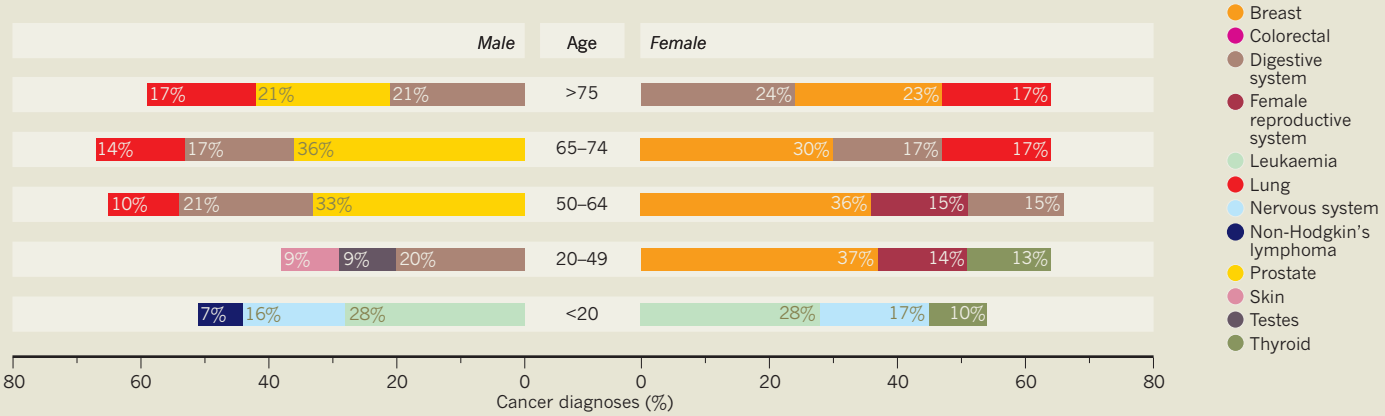
A GLOBAL KILLER

The number of people who die from cancer varies greatly around the world, often because of differences in behaviour and healthcare.



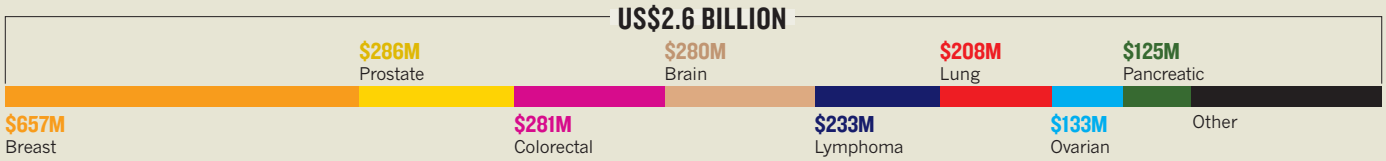
AGE-OLD PROBLEM

A dramatic change happens around the age of 20, when the main cancers being diagnosed in the United States start to shift from mainly leukaemia to predominantly digestive, prostate, lung and breast.



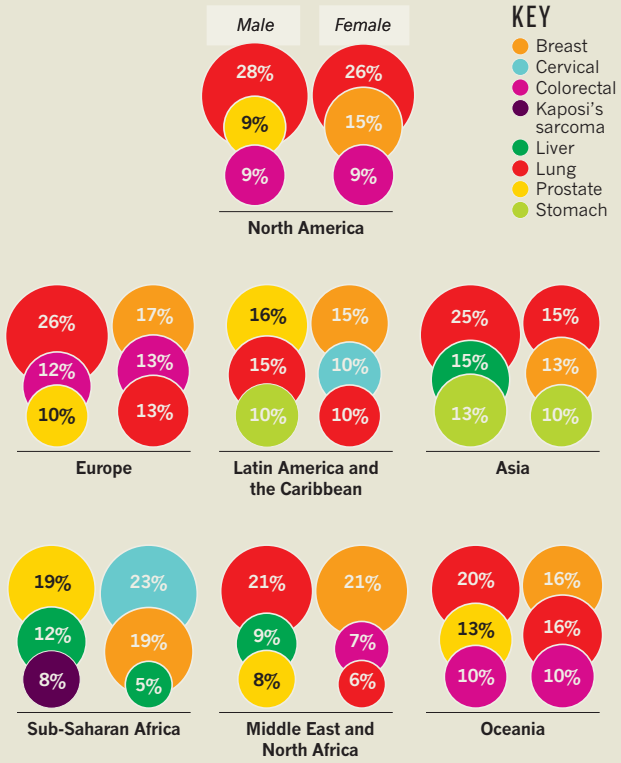
MONEY MATTERS

In 2013, the US National Institutes of Health spent US$2.6 billion on cancer research, and more than one-quarter of that went to breast cancer.



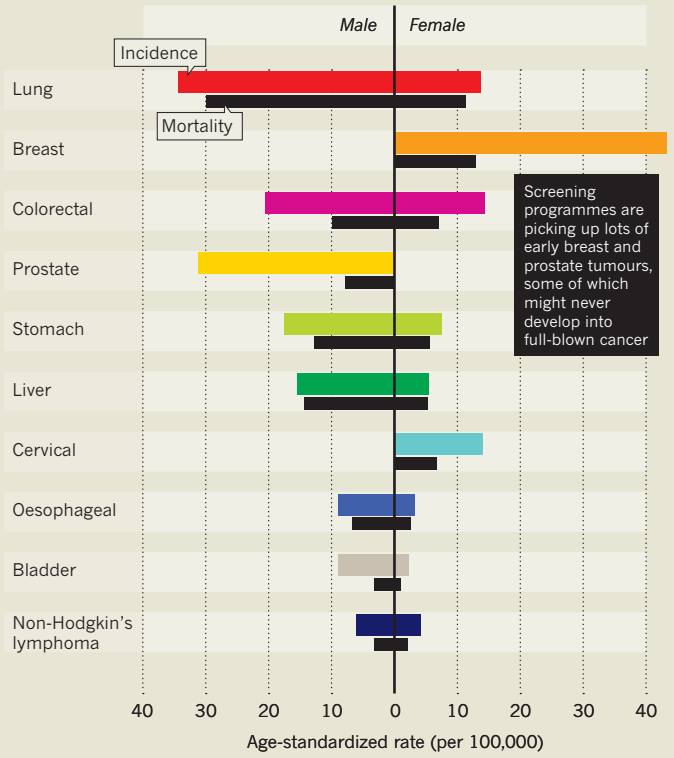
POINTS OF ATTACK

With the exception of sub-Saharan Africa, lung cancer is one of the top three cancer killers in all regions. Breast, colorectal and prostate also feature prominently.



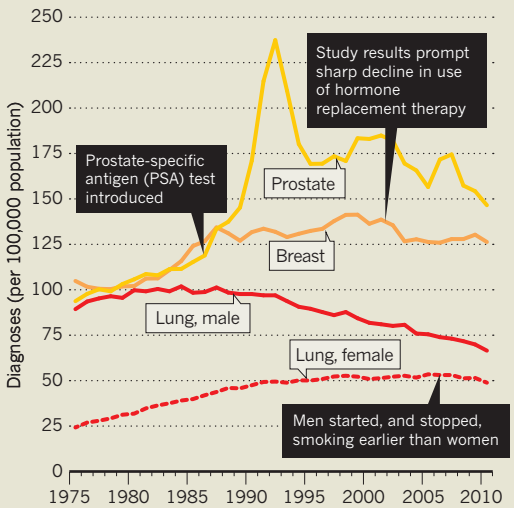
DEADLY DISCREPANCY

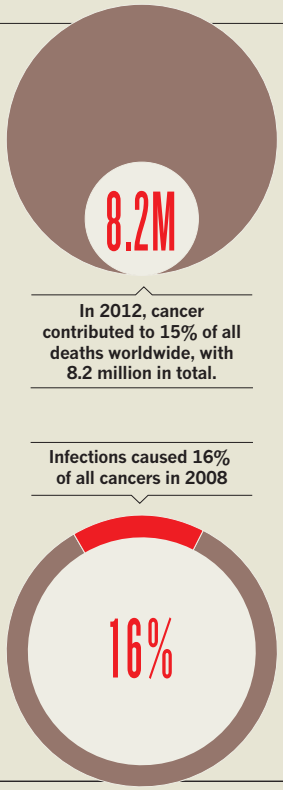
Gaps between diagnoses and mortality are most prominent for breast and prostate cancer.



RATE CHANGES

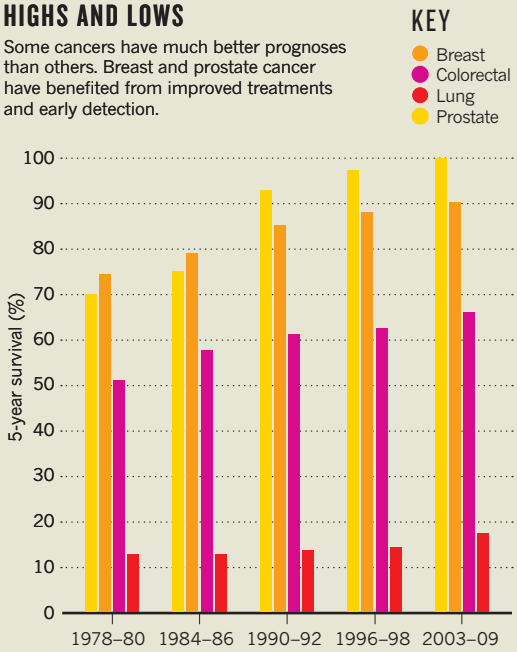
Among other factors, public-health measures have influenced the number of US people being diagnosed with certain cancers.



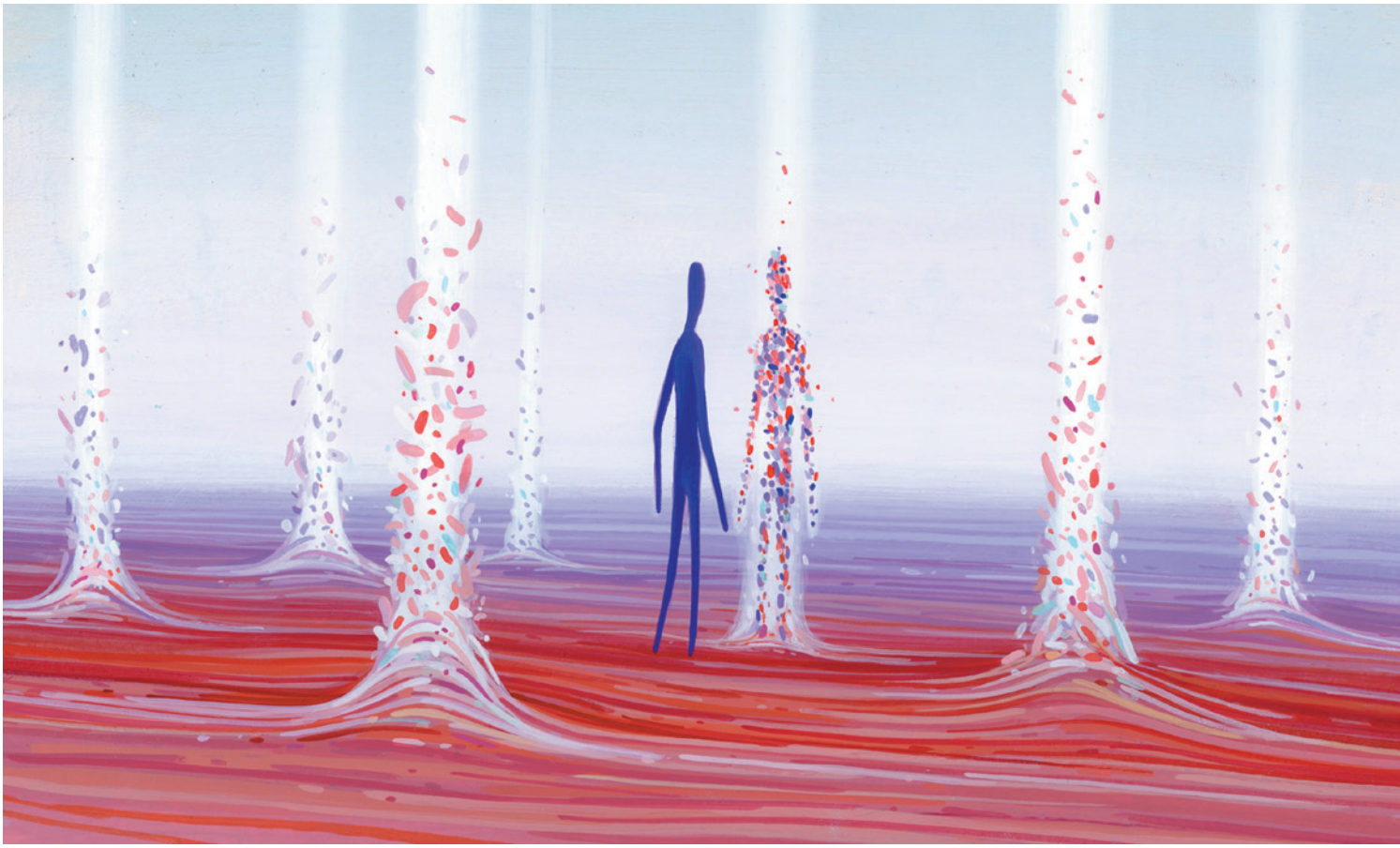


HIGHS AND LOWS

Some cancers have much better prognoses than others. Breast and prostate cancer have benefited from improved treatments and early detection.



A global killer and Points of attack: International Agency for Research on Cancer. Age-old problem, Rate changes and Highs and lows: Surveillance, Epidemiology, and End Results Program. Money matters: National Institutes of Health. Deadly discrepancy: *World Cancer Report 2014*.



BRENDAN MONROE

THERAPY

This time it's personal

Tailoring cancer treatment to individual and evolving tumours is the way of the future, but scientists are still hashing out the details.

BY LAUREN GRAVITZ

Elaine Mardis and her colleagues first encountered 39-year-old Lucy (not her real name) in 2010 at the Genome Institute at Washington University in St Louis, Missouri. Lucy had been referred there after a confusing leukaemia diagnosis. Her doctors thought she had a subtype of the disease called acute promyelocytic leukaemia (APL) — one of the most treatable forms — which usually occurs when parts of chromosomes 15 and 17 get mixed up, or translocated, triggering overproduction of blood-forming cells. But other features of her chromosomes suggested that she might have a much more dangerous form of the disease and therefore need a bone-marrow transplant.

Mardis, co-director of the Genome Institute, is involved in a university initiative to use whole-genome sequencing and other analyses to launch precision attacks against difficult cancers. While her medical colleagues treated Lucy, Mardis sequenced Lucy's genome and that of her cancer and discovered that the leukaemia was indeed caused by a piece of chromosome 15 inserting itself into chromosome 17 (ref. 1). "Our chromosomal analysis indicated that she would respond well to traditional APL therapy," Mardis says. In other words, the treatment she had already received should hold her cancer at bay — and no risky transplant would be needed.

Personalized, 'precision' medicine for cancer is in a difficult time of transition. There are promising stories like Lucy's, wherein the DNA typing of tumours suggests clear approaches to therapy, with improved results for patients. But the field is still limited by many complexities and constraints.

Researchers have learned enough about cancer to know that the way it has been tackled for decades — with cocktails of chemotherapeutic drugs that indiscriminately hit populations of rapidly growing cells — is effective only up to a point. They believe that if they can find the key genetic mutations that drive a particular cancer's growth, they will be able to target the tumour more selectively and with fewer toxic side effects. But they don't yet know enough about which genetic mutations drive a given cancer, let alone how to interrupt the aberrant cellular pathways that result.

MAKE ME A MATCH

Every cancer has a weak spot — a genetic vulnerability that could be exploited by the right drug — and many envision a day when the genome of every cancer will be sequenced, in full or in part, and then paired with an appropriate therapy.

Researchers point to the effectiveness of imatinib (marketed as Gleevec and Glivec) against chronic myelogenous leukaemia (CML) — a rare blood cancer — as perhaps the greatest success in the personalized cancer field so far. CML is most often caused by an abnormal gene rearrangement in which pieces of two chromosomes switch places with each other. Assessing whether a patient is a candidate for the drug requires the analysis of a small group of genes in what is referred to as a gene panel.

"In the 1980s, unless you got a bone-marrow transplant, the disease was an absolute death sentence in four to six years," says Razelle Kurzrock, director of the Center for Personalized Cancer Therapy at the University of California, San Diego. "Today, average survival is more than 20 years. And because the average age at diagnosis is 60, it's almost a normal life expectancy." That success comes at a price: in 2012, a year's worth of the therapy cost US$92,000.

Imatinib's success has not been easy to duplicate. Every tumour has a unique set of genetic mutations — tumours are commonly likened to snowflakes, each is slightly different from the next. And this heterogeneity, which is found even between cells in a single tumour, means that matching a patient with the appropriate therapy can be a complex proposition.

Vulnerabilities such as the one that imatinib capitalizes on are known as driver oncogenes, genetic changes that generate the proteins driving a cancer's growth. Disabling these proteins should, at least in theory, beat back the disease. The number of driver oncogenes seems to be limited — perhaps as few as 200–300 common ones, says Robert Nussbaum, a medical geneticist at the University of California, San Francisco. Understanding how to disable the common driver oncogenes should therefore enable the treatment of a large number of cancers. "First, we have to know what the genes are and how are they mutating. Then, the second challenge is developing drugs that target these abnormally activated proteins," Nussbaum says.

Such an approach means that oncologists are no longer limited to treating cancer on the basis of the organ in which it first appeared. "The whole idea of starting to classify tumours by their mutations and expression profile as opposed to the way they look under the microscope is another branch of this precision oncology that's developing," Nussbaum adds. A case in point: imatinib is not only good at keeping CML in check, it works for certain gastrointestinal cancers and other tumours as well.

"For the first time, we have a landscape of all the frequent mutations that occur in every single major cancer type," says José Baselga, a cancer biologist at the Memorial Sloan-Kettering Cancer Center in New York. "We know which are the frequent mutations that occur in breast cancer, we know which occur in all forms of thyroid cancer, leukaemia, lymphoma, CML — you name it."

Baselga and his colleagues are using this information to design clinical trials that group patients by genotype rather than by a cancer's organ of origin. For example, mutations in the gene *BRAF* can cause the protein it encodes to become oncogenic. The team has been testing a drug called vemurafenib (Zelboraf), which is effective against melanomas that contain a mutation in the BRAF protein known as BRAF(V600E), in patients with other types of cancer who test positive for the same mutation.

"We are beginning to see responses in tumour types that we would have never guessed," Baselga says. "We have very high responses in histiocytosis, hairy-cell leukaemia and some forms of thyroid cancer."

CATCH 22

Drugs with precise molecular targets such as the one that Baselga is testing can be dazzlingly effective in the short term. But that brilliance is dimmed by a massive cloud: because cancer is a continually evolving disease, such therapies rarely retain potency in the long term. Adapting mutations eventually allow cancer cells to grow back in treatment-resistant forms. "Tumours evolve for a living," Nussbaum says. "When you treat them with a targeted therapy, it's a perfect Darwinian system for selecting exactly the cells you don't want."

Because cancer cells evolve ways to survive when one oncogenic pathway is blocked, researchers are seeking to identify not just one but all of the potentially malignant pathways so as to hit them simultaneously — and curtail the ability of tumours to evolve resistance. "Mutations are occurring all the time," Kurzrock says. "Targeting just one abnormality means you're constantly chasing your tail."

Kurzrock ran into this problem in a study in which her group used a gene panel from a company in Cambridge, Massachusetts, called Foundation Medicine (see 'Testing times') to test 75 women with advanced breast cancer. Although the patients each had, on average, five or six malignancy-linked mutations, none had the same combination.

"Not only do we have next-generation sequencing-based methods, we also have this incredible growth in our general knowledge."

Why is it, then, Kurzrock asks, that we have been trying to fit these differently shaped pegs into the same round hole? Instead, she says, "we should take a patient and ask: 'What cocktail of drugs does this particular patient need based on their particular profile?'".

Another problem with the targeted approach is that therapies are typically tested only on patients with advanced cancers, which are much harder to treat than those in their early stages. Trying out drugs earlier in the course of a disease, when the cancer is more likely to be driven by just one or two key mutations, would require a major shift in the clinical trial system (see page S55).

But perhaps the biggest obstacle in targeting the products of mutated genes is that so many of the causative mutations result not in something's presence but in its absence. Most of the driver oncogenes have what is known as loss-of-function mutations — changes that disable the genes or proteins normally

TESTING TIMES

Biotechnology companies are developing sophisticated ways to match patients to therapies — and even determine whether therapy is necessary. Here are some of the most prominent.

Company	Goal	Technology	Developmental stage
Genomic Health, Redwood City, California	Risk assessment	Assesses molecular markers in tumours. Predicts whether chemotherapy will be beneficial, as well as likelihood of recurrence.	Launched onco*type* DX test for breast cancer in 2004. Tests now also available for colon and prostate cancers.
Epic Sciences, San Diego, California	Diagnosis and monitoring	Isolates tumour cells circulating in blood, and tests them for receptors and enzymes that indicate effectiveness of therapeutics.	Partnering with pharmaceutical and biotechnology companies and cancer centres. Tests are currently in clinical trials and not yet commercially available.
Foundation Medicine, Cambridge, Massachusetts	Match tumour to drug	Screens biopsies for alterations in 236 cancer-related genes for solid tumours and 405 genes for haematological cancers. Then matches mutations to drugs that are either approved by the US Food and Drug Administration or in clinical trials.	Two clinical products are available to oncologists: FoundationOne for solid tumours, launched in 2012; and FoundationOne Heme for haematological cancers, launched in 2013.
Qiagen, Hillden, Germany	Targeting appropriate patient subgroups	Uses the polymerase chain reaction analysis to detect mutations in epidermal growth factor receptor and determine whether the drug afatinib is the appropriate treatment.	Approved by US Food and Drug Administration for testing in conjunction with afatinib to select metastatic non-small-cell lung cancer patients for first-line treatment with afatinib.
Trovagene, San Diego, California	Monitoring	Analyses cell-free cancer DNA in urine to detect mutations and monitor disease progression, recurrence and therapeutic response.	Urine testing available for *KRAS* and *BRAF* mutations, which are predictive of response to colon cancer and melanoma therapies.

responsible for preventing cancerous cells from growing out of control. "It's one thing to develop a drug that blocks an activated protein. It's quite another to develop a method to compensate for the loss of a tumour-suppressor protein," Nussbaum says. Attacking these kinds of cancers will require a more nuanced approach — one that tinkers with the DNA itself rather than the proteins for which it codes.

TIME IT RIGHT

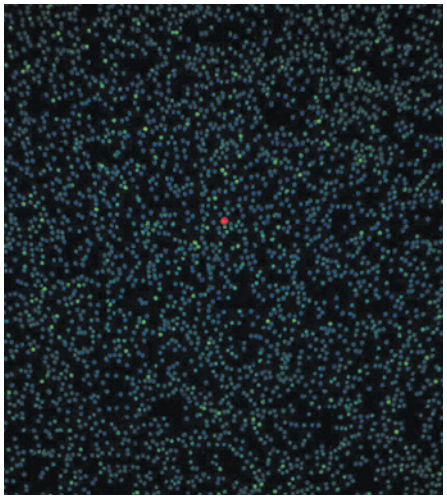
Beyond matching tumour to drug, precision medicine also depends on providing the drugs at the right time — something that requires knowing not only which mutations got a tumour started, but also how the tumour is likely to change. To ensure that therapies that start out personalized remain that way, a doctor needs to know when new oncogenic pathways pop up and when it is time to change course. But repeated biopsies are difficult, and often impossible.

Researchers have therefore been working on non-invasive ways to monitor mutations. Technology is getting good enough to separate out tumour cells or sequenceable scraps of DNA from blood samples, making it possible to do 'fluid' biopsies that provide accurate biomarkers for assessing the disease over time[2]. For instance, Epic Sciences in San Diego, California, in collaboration with cell biologist Peter Kuhn at the nearby Scripps Research Institute, has developed a way to separate tumour cells from blood and assess them for mutations and abnormal protein expression. Others are focusing on bits of DNA that have leaked from dying cells and can be analysed for cancer-driving mutations.

Such methods could one day allow real-time assessment of a patient's tumour make-up. Sarah-Jane Dawson, a molecular biologist and oncologist at the Peter MacCallum Cancer Centre in Melbourne, Australia, studies cell-free DNA — DNA that has escaped from dying cells and is circulating in the blood. She and her colleagues have found that changes in cell-free tumour DNA are detectable, on average, five months before any changes to a patient's cancer are seen in computed tomography (CT) or other scans. "That's not an insignificant amount of time for someone to remain on a therapy they're growing resistant to," Dawson says.

Researchers are also beginning to understand more about how the body fights cancer, and how to take advantage of that. Certain types of immunotherapy — treatments that prompt the body's immune system to detect and attack a tumour — seem to work better if the disease is slightly more advanced: the more mutations a cancer has, the more foreign proteins there are for immune cells to detect.

Tumour cells can be adept at exhausting, or simply turning off, the body's immune response. The best approach to treatment, therefore, might be to combine precision therapy with immunotherapy. First, prune back the cancer with molecularly targeted drugs. Next, use immunotherapy to help the patient's immune system recognize and attack the mutant cells at the first sign of danger[3]. "The vast majority of patients respond to genomically targeted therapies, but with short durability," says James Allison, a cancer immunologist at the University of Texas MD Anderson Cancer Center in Houston. Immunotherapy, he says, is the opposite. "A fraction of patients respond, but with long durability." Combining the approaches, he says, should dramatically improve outcomes.



EPIC SCIENCES

Using just a blood sample, Epic Sciences can detect a single cancer cell (red) among a field of white blood cells (blue and green).

ATTACK FROM WITHIN

Allison has pioneered a class of immunotherapy drugs called checkpoint inhibitors, which can set loose an otherwise-blocked immune system and allow it to break through the defences of certain cancers. Most notably, drugs that inhibit the a protein found on the surface of T cells called PD1, or the PDL1 protein to which it binds, have been shown in numerous clinical trials to be effective against various types of advanced cancers[4]. And Allison's colleague, oncologist and immunologist Padmanee Sharma, is seeking markers that could indicate whether a patient is responding to another checkpoint inhibitor that works against a T-cell surface receptor, CTLA4. One of the markers she has found — inducible co-stimulator (ICOS) — seems to increase when someone is responding to treatment that targets that receptor[5].

Another, still-more personalized form of immunotherapy genetically engineers a patient's immune cells (or those from a matched donor) so that they can recognize and attack cancer cells. The approach, called chimaeric antigen receptor (CAR) T-cell therapy, has produced encouraging results in several small unpublished clinical trials for advanced blood cancers: some patients achieved complete remission.

Ideally, a treatment should be personalized and hit multiple pathways simultaneously. Chemical engineer Mark Davis from the California Institute of Technology in Pasadena and cancer biologist Frank McCormick from the University of California, San Francisco, believe that the answer lies in RNA interference, a technique that uses double-stranded sequences of 'short interfering' RNA (siRNA) to mute specific genes.

Davis has created a way to encapsulate cancer drugs in nanoparticles that have the right size and surface properties to be taken up by tumour cells (see page S58). He is now working with McCormick to infuse these nanoparticles with siRNA. Until now, delivering these fragile molecules to cancer cells had proved nearly impossible, but Davis's nanoparticles provided an elegant solution. The approach has been tested in a phase I clinical trial in solid cancers[6], and the results are now being assessed.

The siRNA method could disable cancer at its very origin by silencing the genes responsible. "In a dream situation, you find a set of genes that affect your tumour, load them up and go," McCormick says. As a patient's tumour evolves, an oncologist can simply swap old siRNAs for new ones. "Once the delivery system works, you could just plug and play different payloads."

Combining a single delivery system with pluggable siRNAs would make drug development faster, cheaper and more routine, and also present less risk to the patient, McCormick says. The multi-siRNA technique has been tested successfully in mice, but human trials using a combination of siRNAs may still be a few years off — such fast-turnaround, individualized therapies pose a challenge for regulatory bodies such as the US Food and Drug Administration.

Cancer is a wily enemy and protects its secrets well. Precision approaches, such as the one Lucy received, are not yet available for most patients. But the rush of research suggests that it is only a matter of time before they are. "Not only do we have next-generation sequencing-based methods," Mardis says, "we also have this incredible growth in our general knowledge."

For Lucy, at least, that knowledge is everything. Four years on, she is still cancer free. ■

Lauren Gravitz *is a freelance science writer based in Los Angeles, California.*

➲ NATURE.COM
For more on immunotherapy approaches, see:
go.nature.com/vdnyet

1. Welch, J. S. *et al. J. Am. Med. Assoc.* **305,** 1577–1584 (2011).
2. Murtaza, M. *et al. Nature* **497,** 108–112 (2013).
3. Vanneman, M. & Dranoff, G. *Nature Rev. Cancer* **12,** 237–251 (2012).
4. Topalian, S. L. *et al. N. Engl J. Med.* **366,** 2443–2454 (2012).
5. Fan, X. Quezada, S. A., Sepulveda, M. A., Sharma, P. & Allison, J. P. *J. Exp. Med.* **211,** 715–25 (2014).
6. Davis, M. E. *et al. Nature* **464,** 1067–1070 (2010).

AMELIE-BENOIST/BSIP/BSIP/CORBIS

With standard treatments being replaced by more personalized ones, trial design needs to change, too.

CLINICAL TRIALS

More trials, fewer tribulations

Clinical studies that group patients according to their molecular profile can make for better and faster drug approval decisions.

BY MICHAEL EISENSTEIN

In December last year, a breast-cancer trial for the experimental drug neratinib captured industry attention — but the buzz was not just about the drug.

What was unusual was the trial itself. Known as I-SPY 2, it assesses multiple drug candidates in parallel, instead of the usual practice of one at a time. The approach is part of a wave of efforts to reform the costly and time-consuming process of drug approval that often fails to take into account the complex realities of cancer biology.

In I-SPY 2, each drug is screened in patients whose tumours have specific molecular profiles. The trial 'learns' as it accumulates data, so rather than randomly assigning new patients to just treatment or control, it uses early results to adjust recruitment. Made by Puma Biotechnology in Los Angeles, California, neratinib was just one of five targeted compounds being tested, and all were designed to selectively block signalling pathways involved in tumour growth.

The standard road to drug approval involves demonstrating safety in phase I, clinical effect in phase II and then a phase III randomized controlled trial (RCT) to confirm whether the experimental treatment provides a statistically meaningful improvement over the current standard of care. RCTs have enabled the discovery of valuable treatments that bolster both survival time and quality of life. "We actually had some outstanding successes early on — like childhood leukaemia, where we saw small improvements from various drugs stack up until the disease turned into something that is usually cured," says Richard Kaplan, a medical oncologist at the UK Medical Research Council's Clinical Trials Unit in London.

Progress against cancer has since slowed down, but many oncologists are hopeful that it is poised to accelerate once more. Thanks to a deeper knowledge of genetics and cell biology, the blunt instrument of cytotoxic chemotherapy — which indiscriminately targets all rapidly dividing cells — is now being supplemented by drugs created for tumours with specific molecular features, or biomarkers. But clinical-study design has not kept pace. Many RCTs still tend to take a broad view, making relatively simple comparisons of drug performance in two roughly identical patient groups. But their failure to account for individual genetics means that they can give rise to misleading results.

Witness the tale of gefitinib, a targeted drug developed by AstraZeneca in London and marketed as Iressa. After showing early promise in some patients with non-small-cell lung cancer (NSCLC), the drug failed in a phase III trial in 2005 (ref. 1). The trial's nearly 1,700 patients had not been selected on the basis of their tumour mutational profile. "The company took the tack of trying to get all of NSCLC," says Donald Berry, a biostatistician at the MD Anderson Cancer Center in Houston, Texas. "It hoped that the benefit in this small subset would drive things." The poor results of the trial led the US Food and Drug Administration (FDA) to put

severe restrictions on who could be prescribed the drug. Later analyses[2], however, revealed that the drug was effective in a specific subset of patients, and gefitinib is now available in Europe to patients with the appropriate mutations.

Critics point to the gefitinib story as a collision between new drugs and old trial design. They assert that conventional randomized trials are too costly, delay the identification of good therapies and mask the benefits of good drugs that work in only a subset of patients.

OPEN ARMS

In one way, however, gefitinib is an example of progress in getting drugs to patients more quickly. It is one of a number of oncology drugs to be approved for use by the FDA through its accelerated approval programme. The programme allows drugs to be marketed if they show strong evidence of clinical effect in a phase II study as long as a subsequent phase III trial is done to confirm the effect. (This is where gefitinib fell down and gained its tight restrictions).

With numerous candidates in their pipelines, pharmaceutical companies must make difficult decisions about how to invest their resources. When many separate trials are done in parallel, they compete for a limited pool of patients. One study[3] showed that filling all pancreatic-cancer trials in the United States in 2011 would have required the participation of 83% of patients with surgically treatable tumours. Yet only about 5% of patients volunteer for trials, according to the American Cancer Society.

Multi-armed adaptive trials such as I-SPY 2 and FOCUS4 — a colorectal cancer trial that started recruitment in January — offer a way to tackle limits on both company resources and the number of available patients. These phase II trials study several markers and drug candidates at once, responding to results by expanding studies for promising treatments and discontinuing them for those that are not showing any effect (see 'Adaptive design').

I-SPY 2 divides patients with breast cancer into ten subgroups on the basis of their tumour's molecular profile. Each subgroup is then divided among the treatment and control groups. Responses to each drug are compared against a single control arm. Future recruitment is not strictly randomized, but rather is informed by incoming trial data. This way, drug–biomarker combinations with early promise are allocated more patients with the same biomarker profile.

"We will need to test a new strategy of customization."

"This is about updating knowledge as you go and modifying your actions on the basis of that knowledge," says Berry, who co-organized the trial with breast-cancer specialist Laura Esserman from the University of California, San Francisco. I-SPY 2 has a second graduate moving on to phase III trials — the drug veliparib from AbbVie in North Chicago, Illinois. Five other compounds are still being tested.

FOCUS4 is recruiting patients to four treatment arms. Unlike in I-SPY 2, patients are assigned to the treatment for which their biomarker profile is thought to be a match. Each treatment arm has its own control group, made up of patients with the same biomarkers. Drugs that perform well in their biomarker-matched group will also be given to individuals whose tumours lack that marker to test for broader effects. A separate chemotherapy-only arm will treat patients who, for whatever reason, cannot participate in the other groups, as well as those who have not responded to the treatment on trial but might benefit from future drugs that target their tumour subtype.

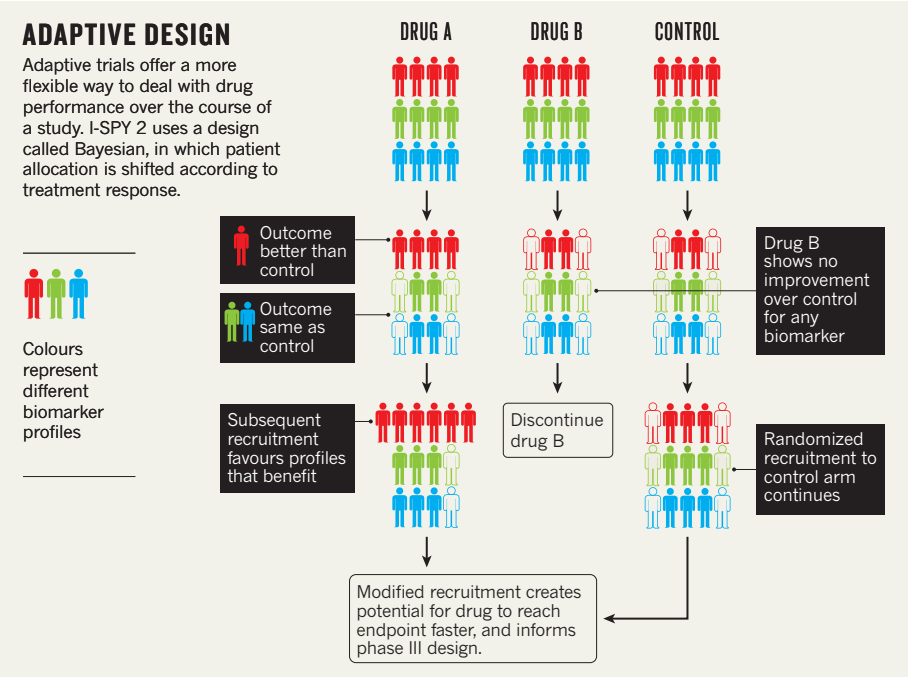
In another shift from business as usual, I-SPY 2 is focusing on initial treatment, rather than limiting itself to patients facing poor prognoses from advanced, metastatic or drug-resistant disease. "Looking at metastatic disease is always first in cancer, and if nothing happens you don't continue," says Berry. "We have to look earlier." Women in I-SPY 2 receive 'neoadjuvant' treatment that is intended to shrink their tumours before they are removed. Trial designers have tended to shy away from early-stage patients who might already be curable with standard treatments, but early-stage tumours often have fewer mutations and are more homogeneous, so could be easier to target.

Such early-stage testing has led to improved outcomes in chronic myelogenous leukaemia (CML), says Razelle Kurzrock, director of the Center for Personalized Cancer Therapy at the University of California, San Diego. There is already an effective targeted drug for CML: imatinib, which Novartis markets as Gleevec or Glivec. When physicians were using this drug only as a last resort, imatinib offered limited returns. But Kurzrock says that when doctors started giving it to patients upon diagnosis, the improvement in performance was dramatic. "The response rate is no longer just 10%," she says. "It's close to 100%."

Encouragingly, the FDA declared in mid-2012 that it would consider accelerated approval for breast-cancer drugs that can eliminate detectable tumour tissue without surgery[4], based in part on data from trials such as I-SPY 2. In September 2013, the agency issued its first such approval for the neoadjuvant use of pertuzumab, which Roche markets as Perjeta.

These trial designs offer greater opportunities for patient participation by creating treatment groups for almost all comers, rather than simply rejecting patients who do not match a single-biomarker criterion. Furthermore, I-SPY 2's sole control arm yields considerable savings relative to the expense of having a control group for each treatment. Both FOCUS4 and I-SPY 2 also offer the potential for even greater cost-cutting by seeking stronger gains from the drugs than those generally sought in clinical trials — typically, a doubling of survival without tumour progression on the treatment drug than on the control. This reduces the number of patients needed to obtain robust phase III data and ensures that only high-performance candidates move forward. "If a drug doesn't meet its prespecified outcome during interim analysis, we're going to close that arm," says Kaplan.

Successful graduation from I-SPY 2 requires a projection that a drug has an 85% chance of succeeding in a 300-patient phase III trial, and treatment arms in FOCUS4 can move seamlessly into phase III if participating companies opt to continue. This is also a feature of the



Lung Cancer Master Protocol, an adaptive trial for squamous-cell lung cancer developed with support from the advocacy group Friends of Cancer Research in Washington DC. The trial is expected to start recruiting soon.

DISEASE REDEFINED

As genomic studies start to provide further information about the mutation profiles of different cancers and as more biomarkers emerge, researchers are re-evaluating cancer classification (see 'Second chance'). A colorectal cancer that shares a mutation with a breast carcinoma may have more in common than two breast carcinomas with different mutations, for example. If this proves to be the case, then these should be the similarities that inform drug testing.

Kurzrock is among many who favour molecular profiling over tissue-based definitions. "If you have a drug that targets a specific abnormality, you would want to look at that abnormality — not whether you're dealing with breast cancer," she says. Evidence to support this model is mounting. For example, although the FDA has approved crizotinib (marketed by Pfizer as Xalkori) for NSCLC, clinical studies suggest that the drug could also be effective for children with aggressive brain tumours that have the same mutation[5].

Such approvals must now be won gradually through trials on different diseases. To speed things up, several companies are pursuing 'basket' trials that test treatments on multiple cancers with common genetic disruptions. GlaxoSmithKline is testing two melanoma drugs, dabrafenib and trametinib, in nine cancers — including brain, thyroid and intestine — that share mutations in the gene *BRAF* that could render them susceptible to these drugs. Rafael Amado, senior vice-president for oncology research and development at the firm, argues that this approach offers hope to patients with rare cancers who might otherwise slip through the cracks. By performing analyses that take data from across tumour groups, even small sets of positive outcomes can become statistically meaningful. As a result, says Amado, "we don't have to run very large randomized trials in these ultra-rare populations".

The US National Cancer Institute is exploring this approach through its Molecular Analysis for Therapy Choice programme, using targeted gene sequencing to match various drugs to people with solid tumours or lymphomas whose disease has progressed on existing treatments. "We'll have about 20 arms to start with, targeting the usual suspect mutations that you might find in cancer," says Barbara Conley, associate director of the institute's cancer-diagnosis programme. "If we can get 35% or more patients across tumour types to survive six months or more, that's an interesting signal." These are essentially phase II trials looking for indications that could justify a move to phase III, but regulators say that they are willing to formally recognize robust evidence of cross-tumour efficacy. "The FDA could approve a drug based on a molecularly defined population rather than a disease-site-specific indication," says Richard Pazdur, director of the agency's Office of Oncology and Hematology Products.

➲ NATURE.COM
For more on clinical-trial design, see:
go.nature.com/do8cae

Additional complexity could confound this broad biomarker-informed research, however. For instance, *BRAF* inhibitors that work in some melanomas are ineffective in colorectal tumours with the same mutations. But Kurzrock maintains that universal effectiveness is an unrealistic expectation. "If you look at drugs that were approved for lung cancer, the response rates were usually in the range of 15–20% of patients," she says. "We cannot expect 100% of patients treated on the basis of a genomic classification to respond."

Indeed, tumours often contain multiple mutations that might drive drug resistance — a common roadblock for targeted agents suggesting that each patient's cancer may require a specialized cocktail of agents. "We will need to test a new strategy of customization per patient and patient-centric care," says Kurzrock, "rather than just the old way of testing a drug or combination of drugs."

This complexity will mean a steep learning curve for researchers and oncologists. Berry believes that oncology will ultimately undergo a broad transformation — approaching drug testing as an opportunity to gain insight into the disease rather than merely validate existing hypotheses. "The future really is combining clinical practice and clinical trials, and having a notion of both learning and confirming in the trial," he says. "It will mean a completely different regulatory perspective and an entirely different business model for companies." ■

Michael Eisenstein *is a freelance science writer based in Philadelphia, Pennsylvania.*

SECOND CHANCE

A new perspective for past drug decisions

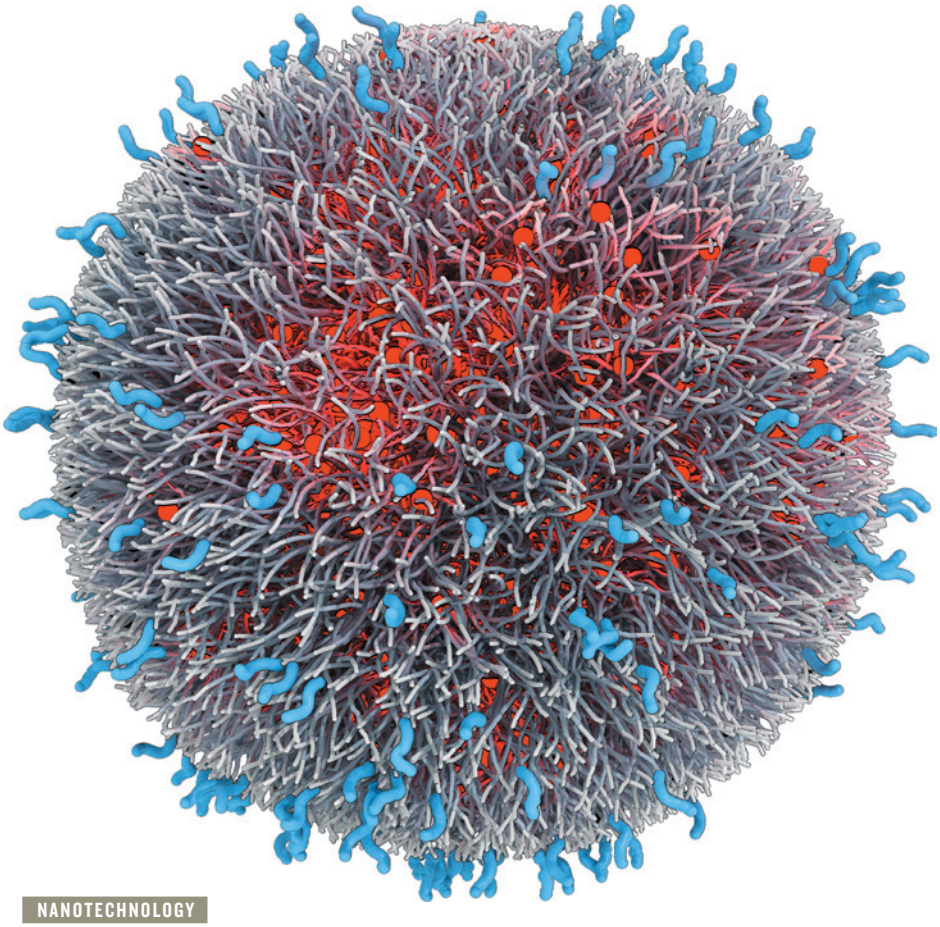
A drug's journey to the clinic doesn't necessarily end with a regulator's decision. As clinical oncologists learn more about the interplay between a patient's response to a drug and his or her genetics, it is becoming clear that some failed drugs might be rehabilitated by looking at instances in which a small subgroup of participants showed significant benefit.

Take everolimus, developed by Novartis. This drug failed a phase II trial for metastatic bladder cancer — but one patient's cancer went away and stayed away for more than two years. Researchers at the Memorial Sloan-Kettering Cancer Center in New York subsequently found that about 8% of people with bladder cancer carry a mutation that makes their tumour susceptible to the drug[6]. The US National Cancer Institute is now examining other reports of rare but significant drug responses through its Exceptional Responders programme, which might see some failed drugs put back into development. "We've looked at some of the phase II trials that didn't get approval over the past few years, and in lots of them as many as 10% of the people had exceptional responses," says Barbara Conley, associate director of the institute's cancer-diagnosis programme.

Conversely, even when a drug is approved, its success can only really be assessed in the real world. "Trial subjects don't have cardiac or renal disease," says Ian Tannock, an oncology researcher at Princess Margaret Cancer Centre in Toronto, Canada. "But then that result is taken out into a community of unselected patients where there is a lot of comorbidity." As a result, small benefits may disappear and rare toxicities may become apparent. Tannock advocates large-scale research to track the performance of drugs after they have been approved. But such observational studies require close collaboration across clinical centres — a process made difficult in the United States, at least, by the lack of a central national cancer registry.

The American Society for Clinical Oncology hopes to rectify this problem with CancerLinQ, which integrates patient treatment and outcome data. "This would allow us to query data for a group of patients with a particular set of molecular characteristics, look at what treatments those patients received and identify the treatment that seemed to work the best in that subset of patients," says Neal Meropol, a medical oncologist at Case Western Reserve University in Cleveland, Ohio, and a member of society's board of directors. "We can now ask questions that simply could not be answered in a standard clinical trial." **M.E.**

1. Thatcher, N. *et al. Lancet* **366,** 1527–1537 (2005).
2. Maemondo, M. *et al. N. Engl. J. Med.* **362,** 2380–2388 (2010).
3. Hoos, W. A. *et al. J. Clin. Oncol.* **31,** 3432–3438 (2013).
4. Prowell, T. M. & Pazdur, R. *N. Engl. J. Med.* **366,** 2438–2441 (2012).
5. Mossé, Y. P. *et al. Lancet. Oncol.* **14,** 472–480 (2013).
6. Iyer, G. *et al. Science* **338,** 221 (2012).



DIGIZYME/MIT

The Accurin nano-drug BIND-014 encases a toxic payload (red) in a layer of biodegradable polymers (grey), and uses molecules on the surface (blue) to target the tumour.

NANOTECHNOLOGY

Deliver on a promise

Effective treatment of cancer requires getting the drugs precisely to the target. Enter the nanoparticle.

BY JESSICA WRIGHT

When Mary Davis nearly died from the drug used to treat her breast cancer, she issued a challenge to her scientist husband: would he switch his research focus to the design of better cancer treatments? As a chemical engineer at the California Institute of Technology in Pasadena, Mark Davis was creating solid catalysts for use in chemical synthesis. But in 1996, after Mary's experience with the highly toxic chemotherapy drug doxorubicin — dubbed the 'red death' — he heeded her plea and went to work.

Davis used his engineering know-how to design an armour-like particle less than one-thousandth the width of a human hair to encase the drug camptothecin. About ten years after his wife's treatment, he watched doctors inject this molecule, now called CRLX101, into the first patient, Ray Natha. "It was the scariest thing I've ever done," Davis recalls. During the six-month safety trial, Natha's pancreatic cancer, which had spread to his lungs, stalled. And the nanoparticle-coated drug seemed to cause fewer side effects than did unprotected drugs.

Building protective coats around toxic molecules could address one of cancer treatment's biggest remaining challenges — how to spare healthy cells when attacking cancerous ones. Chemotherapy drugs kill rapidly dividing cells to prevent the rampant cell growth that results in tumours. But these drugs reach cancer cells through the same network of blood vessels that supplies the whole body. And, as Mary Davis's experience shows, they are just as likely to poison healthy cells as cancerous ones.

Researchers are therefore working on ways to deliver drugs directly to cancerous tissue. Drug delivery is a multi-stage journey: the active agent needs to enter the body, travel through the bloodstream to arrive at the tumour site, penetrate the tumour mass and then gain entry to the cells (see 'Hole in one'). By refining each step, researchers aim to do more than just protect the body from toxic medication. Pioneering drug-delivery systems are designed to transport payloads — from highly toxic molecules to genetic material — that should never travel through the body alone, and to target cancer cells more precisely than is now possible.

ARMED ANTIBODIES

Much of the excitement over developments in cancer therapy has focused on drugs that target cancer-specific biological pathways, says Steven Libutti, a cancer surgeon at the Albert Einstein College of Medicine in New York. "But this method may not deliver the promise that we hoped, because the tumours themselves evade that strategy," he says. The alternative is to use potent drugs that are toxic to all cells, but to corral these in a benign 'Trojan horse' until they reach the tumour, he says.

One of the simplest ways to do this is to arm cancer-seeking proteins with a cell-killing drug. In the late 1990s, drug companies developed antibodies that bind to the surface of certain types of cancer cell. Fusing these proteins to drugs with a stable chemical linker yields a potent combination: the antibodies encourage uptake of the drug into cancer cells and the linker keeps the drug from working until it gets inside. "It's a simple idea," says John Lambert, chief scientific officer at ImmunoGen in Waltham, Massachusetts, which develops and licenses the technology that links the antibody to the drug. "But it has taken a long time to put all the pieces together."

Only two antibody–drug conjugate (ADC) therapies are on the market. The first, called brentuximab vedotin (marketed by Seattle Genetics in Washington as Adcetris), was approved by the US Food and Drug Administration (FDA) in 2011 to treat some types of lymphoma, cancers of the lymph system, that had not responded to previous treatment. The second, trastuzumab emtansine (marketed by Genentech of South San Francisco, California, as Kadcycla), was approved in 2013 as a treatment for late-stage breast cancer after treatment with conventional chemotherapeutics.

Trastuzumab emtansine fuses emtansine, a toxic chemotherapeutic, to antibodies that bind to a protein receptor called HER2, which is overproduced by about 20% of breast cancers. In a phase III trial that finished in 2012, the nearly 500 women who took the drug lived about

➲ NATURE.COM
For more on nanoparticle drug carriers, see:
go.nature.com/qoruip

five months longer and had fewer side effects than did those on the standard treatment[1]. All women in the study were in advanced stages of the disease, but the drug is now in clinical testing as a first-line treatment thanks to the positive results from the trial, says Lambert. Several similar drugs are now in advanced clinical trials, and ImmunoGen is trying to link emtansine to antibodies that target other cancers, from lymphoma to lung and ovarian cancers.

COATED NANOPARTICLES

Like conventional monoclonal antibodies, ADCs need to make their own way to the tumour site. "The molecules you inject have no idea where the cancer cell is," says Lambert. "If they pass by the cancer cell they can attach, but most go elsewhere."

One approach is to add a coat around chemotherapy drugs. The resulting nanoparticles, which range from about 20 to 100 nanometres in diameter, are too large to escape most blood vessels. But they do find their way out of the leakier ones hastily built by a rapidly growing tumour. As a result, they are thought to accumulate preferentially at the tumour site. However, this phenomenon has been studied mostly in animal models and not in people, says Rudolph Juliano, a pharmacologist at the University of North Carolina at Chapel Hill.

Nanoparticles can carry a stronger payload than can antibodies, encasing thousands of drugs in a single molecule. Unfortunately, however, they can also accumulate in the liver and the spleen, where they provide no therapeutic benefit and can cause side effects. To minimize unwanted effects, researchers coat the particles with a layer of polyethylene glycol, which mimics water and effectively hides the drug from the liver cells that detect and engulf intruders. But minimizing liver uptake is still an important part of nanoparticle design, says Juliano.

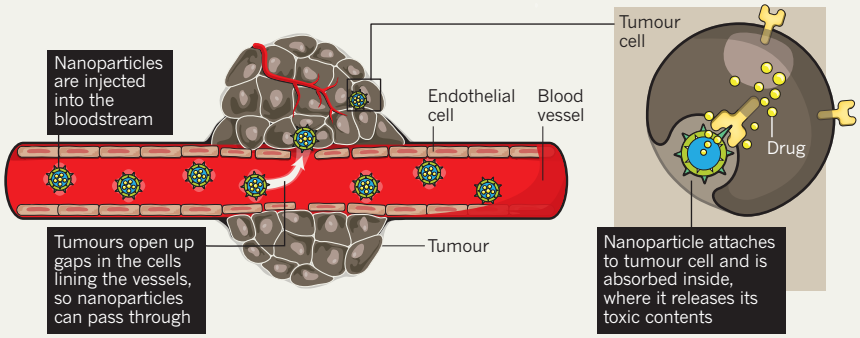
The first nanoparticles to be developed for drug delivery coated the active agent with lipids. The first drug of this type to be approved was Doxil, in 1995. Doxil carries doxorubicin and is used to treat Kaposi's sarcoma and other solid tumours, including breast and ovarian cancer.

According to the US National Cancer Institute, six such nanoparticles are currently approved for use on the market worldwide. So far, they seem to improve safety — Doxil does not end up in the heart, where doxorubicin causes toxicity — but not efficacy. As a result, some researchers have questioned whether nanotechnology is worth the high price tag that accompanies its production: it can cost ten times more than conventional treatment. "New nanoparticle-based drug delivery will be expensive and it has to be justified by improved therapeutic outcomes," says Juliano. "We're still at too early a stage to ascertain that."

Still, nanoparticle drug delivery can have a dramatic and worthwhile effect, says Yun Yen, an oncologist at the City of Hope cancer centre in Duarte, California, who administered CRLX101 (developed by Cerulean Pharma in Cambridge, Massachusetts) to Natha. "It's quite amazing when you see a patient and you're expecting their blood count to drop and you're expecting them to be nauseous, but they do so well," he says.

HOLE IN ONE

Drug-carrying nanoparticles exploit the increased permeability of tumour blood vessels to gain entry to the cell.



ADAPTED FROM D. PEER *ET AL. NATURE NANOTECHNOL.* **2,** 751–760 (2007)

ENGINEERED FOR DELIVERY

Nanotechnology has yet to achieve its full potential because it has so far been used only to ferry drugs intended to be administered through conventional methods, says Omid Farokhzad, a physician-scientist at Harvard Medical School in Boston, Massachusetts, and founder of three nanotechnology-based biotech companies in Massachusetts. One of Farokhzad's companies, Blend Therapeutics in Watertown, is working to engineer drugs specifically for use in nanoparticles.

Farokhzad's other two companies — BIND Therapeutics of Cambridge and Selecta Biosciences of Watertown — use technology that engineers a long polymeric string that spontaneously folds to form a particle. The polymers are interspersed with targeted ligands designed to link the particles to cancer cells. The self-assembly makes it easier for scientists to reproduce the molecule in different batches — a key advantage for translating the technology into the clinic. In a 2012 study, Farokhzad and his colleagues screened 100 polymers that incorporate a molecule that binds to a prostate-specific membrane antigen (PSMA), which is displayed on the surface of most prostate tumours[2]. The particles enclose a chemotherapeutic, docetaxel. The most promising of the nanoparticles, BIND-014, is in phase II trials for treatment of lung and prostate cancer.

"We've built a truck and we have a GPS system. You can load anything on the truck."

Some researchers are branching out from conventional chemotherapy and using nanoparticles to deliver small pieces of RNA to cancer cells, where they decrease expression of certain genes in a method called RNA interference. Davis says that the use of nanoparticles to deliver RNA is promising because it allows researchers to reach multiple genes, and thus pathways, in one hit. He developed the first RNA-carrying nanoparticle to enter clinical trials for cancer. The particle, called CALAA-01, targets the gene *RRM2*, which is involved in cell division and uses molecules that bind to the transferrin receptor, which is highly expressed on cancer cells, to gain access to the cell interior[3]. CALAA-01 entered phase I clinical trials to treat melanoma, but the follow-up trial was halted for reasons that have not been made clear.

Other efforts aim to bring drugs directly to the tumour. Sylvain Martel, a biomedical engineer at Montreal Polytechnic in Canada, is building magnetic particles that researchers will be able to guide to the tumour. These particles use the same contrast agent as in magnetic resonance imaging (MRI), so when a patient is in an MRI scanner researchers can use a strong magnetic field on top of the tumour to guide drugs to the correct site. Martel and his colleagues have tested the method in pigs and aim to try it in people. The magnetic particles can enclose other targeting particles, Martel says. "We've built a truck and we have a GPS system," he says. "You can load anything into the truck."

This vision of a researcher driving a drug directly to the site of a tumour is far from the whole-body onslaught that Mary Davis experienced nearly 20 years ago. Today, the drug she encouraged her husband to develop is nearing the end of phase II clinical trials and has been used to treat hundreds of patients. Future therapies may strike even more efficiently, with fewer side effects, says Mark Davis. "We need to think about ways of driving down cancer so people can have a reasonable life without all the added side effects and toxicities." ■

Jessica Wright *is a freelance science writer based in New York City.*

1. Verma, S. *et al. N. Engl. J. Med.* **367,** 1783–1791 (2012).
2. Hrkach, J. *et al. Sci. Transl. Med.* **4,** 128ra39 (2012).
3. Davis, M. E. *et al. Nature* **464,** 1067–1070 (2010).

PHOTOTAKE/ALAMY

The naked mole rat has been extensively studied, but no cancer has ever been spotted in this species.

COMPARATIVE BIOLOGY

Naked ambition

A subterranean species that seems to be cancer-proof is providing promising clues on how we might prevent the disease in humans.

BY SARAH DEWEERDT

There is a lot not to envy about the life of the naked mole rat: imagine passing your days in a stuffy, pitch-black system of tunnels two or three metres underground with 100 of your closest relatives. But there is one thing that humans might covet: as far as anyone knows, the animal never gets cancer.

Native to the Horn of Africa, this small rodent (*Heterocephalus glaber*) is neither a mole nor a rat; it is actually more closely related to porcupines and guinea pigs. The animal's pale-pink, wrinkled skin is nearly hairless, the better to slip through those narrow burrows. But there is yet another more compelling fact: in all the thousands of naked mole rats that have lived and died in research labs and zoos over the past several decades, not a single instance of spontaneous cancer has been recorded[1].

So far, the animal provides little more than a footnote to the vast body of cancer literature based largely on studies of laboratory mice. But the species has a few fierce advocates in the scientific community, who say that to truly defeat cancer we need to pay a lot more attention to naked mole rats and species like them.

"If we want to learn what naturally occurring resistance mechanisms protect from cancer, we may not find them in mice because mice are even more prone to cancer than humans," says biologist Vera Gorbunova, who co-leads naked-mole-rat studies at the University of Rochester in New York. In some strains of mice, for example, cancer kills 90% of animals.

In other words, although mice are an excellent model for cancer development, progression and treatment, naked mole rats may be better for prevention. "We have to study species that are more resistant," Gorbunova explains.

Researchers have found it tough to induce cancer in naked mole rats. Working in culture, they have infected cells from the creatures with a genetically engineered virus that contains a pair of oncogenes, or cancer-promoting genes, that reliably turns mouse cells malignant[2,3]. "This common oncogenic cocktail had no effect on the naked-mole-rat cells," says Rochelle Buffenstein, a physiologist at the University of Texas, San Antonio, and a pioneer of naked-mole-rat research. "They did not become tumorigenic, they didn't rapidly proliferate, they didn't invade tissues."

Naked-mole-rat cells also seem to be highly sensitive to their neighbours. Normally when cells are grown in culture dishes, they stop proliferating when they touch neighbouring cells; the end result is a smooth, uniform layer covering the surface of the dish. This property, called contact inhibition, is absent in cancer cells. But Gorbunova's team observed that naked-mole-rat cells stop proliferating after just a few cell-to-cell contacts, rather than when all the spaces on the culture plate have been filled in.

The researchers dubbed this hypersensitivity early contact inhibition, and found that it

➲ NATURE.COM
To read more about cancer-proofing in the mole rat, see:
go.nature.com/ygk8gg

is regulated by two genes: *p16* and *p27* (ref. 2). If *p16* is disabled, they showed, *p27* stops cell growth — but at higher cell densities. In humans and mice, *p27* is the main player in contact inhibition, with *p16* taking a minor supporting role. But in naked mole rats, the two genes have become decoupled, resulting in two layers of protection against runaway growth.

In trying to understand what was activating *p16*, Gorbunova's team noticed that naked-mole-rat cells were secreting something into the growth medium that was making it viscous. "We spent some time trying to isolate the goo, and identify what kind of chemical it is," Gorbunova says.

The goo turned out to be hyaluronan, a long polymer of sugars that occupies the spaces between cells in the skin and connective tissues of vertebrates. But in naked mole rats, hyaluronan production is in overdrive: the animals make a large quantity of an unusually large type of the polymer, and break it down more slowly than do other species[4].

That extra dollop of hyaluronan seems to be key to early contact inhibition, because adding enzymes that degrade it to the culture dish blocks the phenomenon[4]. So does blocking CD44, a receptor found on the surface of cells that binds to hyaluronan. How *p16* and CD44 are functionally connected is unknown, Gorbunova says. "We know the beginning point and the end point of the pathway, but we don't understand what's in between."

Others disagree with this interpretation of the results. "We don't see early contact inhibition in our lab," says Buffenstein, who uses a different protocol for growing naked-mole-rat cells. Under optimal culture conditions, she says, the cells will grow to cover a culture dish.

WEIRD AND WONDERFUL

But the cells behave unusually in other respects, Buffenstein has found. They can survive remarkably high doses of heavy metals and carcinogens, although such agents do stop the cells proliferating. By contrast, a high proportion of mouse cells die at low doses — and those that survive continue to proliferate. These multiplying, damaged cells may be the ultimate source of cancers[5].

In a sense, these results parallel Gorbunova's, Buffenstein believes. "We think it's the same mechanism," she says. "When things aren't exactly the way the cell thinks they should be, the cells just sit tight and stop proliferating."

The recognition that something isn't right probably involves *p53*, a tumour-suppressor gene found in many species, including humans. In most species, *p53* is activated only when cells are stressed, but naked-mole-rat cells produce high levels of the protein even under normal conditions. They also express high levels of another protein — nrf2, a master regulator of hundreds of cell-protection genes[6].

All of this suggests that naked mole rats rely on many of the same cancer-protection mechanisms as do humans, just kicked into high gear. "They've upregulated the system to really be extra careful about changes in the cell and when to replicate and when not to," Buffenstein says.

Applying the model of the naked mole rat to human cancer resistance is unlikely to be as simple as upregulating existing anticancer genes. Too much activity of *p16*, for example, can cause cell ageing and death. "My bet would be to try to somehow manipulate turnover of hyaluronan," Gorbunova says. After all, she points out, hyaluronan is already used as a cosmetic treatment for wrinkles, so it may be a relatively feasible target.

Further research may pinpoint other resistance mechanisms in the naked mole rat. "Perhaps many pathways are involved," says Vadim Gladyshev, a cancer biologist at Harvard Medical School in Boston, Massachusetts, who helped to sequence the animal's genome in 2011. "The problem," he adds, "is that the naked mole rat is quite distant from other organisms with completely sequenced genomes." For precisely this reason, he is now sequencing the genome of the Damaraland mole rat (*Fukomys damarensis*). This close cousin of the naked mole rat also lives underground in groups but does develop cancer, so differences between the species could help to identify genomic regions relevant to resistance.

> ***"To some degree we still lack the molecular tools to study them."***

Another challenge in working with such an unexplored model is that "to some degree we still lack the molecular tools to study them", says João Pedro de Magalhães at the University of Liverpool, UK, who specializes in the genomics of ageing. But DNA sequencing and gene-expression technologies can be applied across species, so his lab is surveying how gene expression changes in cells from mice, rats and naked mole rats after exposure to DNA-damaging chemicals. Such patterns may shed more light on exactly how damaged naked-mole-rat cells decide when to stop proliferating.

The naked mole rat is not the only cancer-resistant animal. Scientists are increasingly focusing their attention on the blind mole rat (*Spalax* spp.). These are furry brown torpedoes of rodents that are agricultural pests in the Mediterranean region. "In none of the *Spalax* we've raised, thousands of individuals, have we ever found any cancer," says Eviatar Nevo, founder of the Institute of Evolution at the University of Haifa in Israel.

Despite their similar names, blind mole rats are not closely related to naked mole rats (they are more closely related to rats and mice). The two animals evolved their resistance independently, making for intriguing similarities and differences in the underlying mechanisms.

For example, blind mole rats produce hyaluronan in a form similar to that of naked mole rats, but their cells do not show early contact inhibition when grown in the lab. However, the contact inhibition they do show is not normal: instead of filling a culture dish and then holding steady, blind-mole-rat cells undergo mass cell death when they reach high densities[7].

But the species have in common a *p53* tumour-suppressor gene that functions in an unusual way. In the blind mole rat, this gene has a sequence almost identical to a mutated form of *p53* found in many human cancers, and it seems to encourage cells to stop proliferating rather than to self-destruct[8].

DIRECT ATTACK

Researchers at the University of Haifa have also discovered a tantalizing clue that anticancer mechanisms in the blind mole rat may involve selective destruction of malignant cells. When placed in the same laboratory dish as cells derived from human breast or liver cancers, blind-mole-rat cells reduce survival of the cancer cells[9] — suggesting that they secrete an anti-proliferation factor into the medium.

Many researchers believe that these species' cancer resistance is related to their ability to tolerate the oxygen-poor environment of underground burrows, which is similar to the oxygen-poor environment experienced by cells in rapidly growing tumours. "The idea that an animal living in a cancer-like environment would be able to generate adaptations against cancer is very, very natural," Nevo says. "There are 200 species of subterranean mammal," he adds, suggesting that additional anticancer mechanisms may be found by studying this group of species more thoroughly.

Other researchers are looking more broadly across the animal kingdom, plumbing the genomes of long-lived, cancer-resistant species such as Brandt's bat (*Myotis brandtii*) and the bowhead whale (*Balaena mysticetus*). "Harnessing the power of natural selection to increase our knowledge of cancer and hopefully develop therapies against cancer or for cancer prevention is a very unexplored area," says de Magalhães. The end of cancer may yet prove to be underground — or in the seas or in the air. ■

Sarah Deweerdt *is a freelance science writer in Seattle, Washington.*

1. Delaney, M. A., Nagy, L., Kinsel, M. J. & Treuting, P. M. *Vet. Pathol.* **50,** 607–621 (2013).
2. Seluanov, A. *et al. Proc. Natl Acad. Sci. USA* **106,** 19352–19357 (2009).
3. Liang, S., Mele, J., Wu, Y., Buffenstein, R. & Hornsby, P. J. *Aging Cell* **9,** 626–635 (2010).
4. Tian, X. *et al. Nature* **499,** 346–349 (2013).
5. Lewis, K. N., Mele, J., Hornsby, P. J. & Buffenstein, R. *Gerontology* **58,** 453–462 (2012).
6. Lewis, K. N., Mele, J., Hayes, J. D. & Buffenstein, R. *Integr. Comp. Biol.* **50,** 829–843 (2010).
7. Gorbunova, V. *et al. Proc. Natl Acad. Sci. USA* **109,** 19392–19396 (2012).
8. Ashur-Fabian, O. *et al. Proc. Natl Acad. Sci. USA* **101,** 12236–12241 (2004).
9. Manov, I. *et al. BMC Biol.* **11,** 91 (2013).

Shanghai is one of 74 cities that has yet to meet the air-quality standards set by the Chinese government.

IMAGINECHINA/CORBIS

PREVENTION

Air of danger

Carcinogens are all around us, so scientists are broadening their ideas of environmental risk.

BY REBECCA KESSLER

In November last year, an eight-year-old girl became China's youngest person to get lung cancer. The cause, according to her doctor, was fine particulate matter that accumulated in her lungs and led to malignant changes in her cells. Air pollution has been enveloping Chinese cities in smog, periodically closing schools and businesses, and drastically reducing visibility.

A month before the girl's diagnosis, outdoor air pollution and one of its main constituents, particulate matter, were declared carcinogenic to humans by the International Agency for Research on Cancer (IARC). The announcement capped a decade-long review that examined the cancer-causing potential of several airborne pollutants, including dusts, solvents and metals emitted by vehicles, industry, farms, homes and natural sources.

Scientists have suspected since the 1940s that air pollution causes lung cancer, but it has taken seven decades of research to establish the connection. During that time it became clear that smoking causes most lung-cancer deaths (70% globally) and that air pollution kills more people through cardiovascular disease than through cancer. Nevertheless, air pollution's cancer toll adds up. Researchers blamed it for 223,000 lung-cancer deaths in 2010, nearly 15% of all such deaths. The IARC also noted evidence linking air pollution to bladder cancer[1].

"Everybody is exposed to it," says Aaron Cohen, an epidemiologist at the Health Effects Institute in Boston, Massachusetts, a research organization funded by the US government and the motor-vehicle industry. "You can't avoid breathing the air no matter who you are."

Scientists are making progress in understanding the effects of air pollution and other environmental carcinogens. They are learning that certain chemicals may increase cancer risk at lower-than-expected doses, that people may be particularly vulnerable during certain periods of their lives and that the consequences of exposure may cause ripples for generations. Researchers are also looking into environmental triggers that can lead to the onset of particular cancers. Although it is the quest for a cancer cure that draws the most funding and talent, a small but vocal chorus is calling for more attention to environmental carcinogens, in the hope that reducing exposure to them will help to keep the disease from starting.

The IARC has evaluated a total of 970 natural and artificial agents and identified 464 as having some level of carcinogenicity to humans. With some overlap, a catalogue by the US National Toxicology Program lists 240 substances as 'known' or 'reasonably anticipated' human carcinogens. Some cancer-causing agents occur naturally, such as aflatoxins — poisonous compounds produced by moulds that grow in nuts, seeds and legumes. Others are man-made, such as ionizing radiation from medical imaging and various commercial chemicals. Yet of the 80,000 chemicals in commerce, only a tiny fraction has been tested for carcinogenicity.

The proportion of cancers attributable to environmental carcinogens is subject to debate. The most widely cited estimate, made in 1981, attributes 2% of US cancer deaths to pollution

➲ NATURE.COM
To read more about pollution's link to cancer, see:
go.nature.com/srjaod

and 4% to occupational exposures[2]. Those figures are dwarfed by the numbers for smoking and diet, which claimed 30% and 35% of the burden, respectively. But a 2010 report by the US President's Cancer Panel assailed the 1981 estimate and stated that "the true burden of environmentally induced cancer has been grossly underestimated"[3]. And a global estimate by the World Health Organization (WHO) is much higher: 19% of the world's cancers — and 1.3 million deaths annually — are attributed to environmental and occupational factors.

Even so, many researchers, charitable foundations and government agencies continue to underestimate the importance of environmental carcinogens, says Richard Clapp, an epidemiologist at the University of Massachusetts Lowell. "There's more to it than is generally given credence that environmental exposures are causing a portion of our cancer burden and that we can do something about it," he says.

Designating a substance as a carcinogen and making recommendations about its use can be controversial. For example, in 2011, the US National Toxicology Program listed formaldehyde, a chemical common in building materials and household products, as a carcinogen, and styrene, which is used to make plastics and rubber, as reasonably anticipated to be a human carcinogen[4]. Despite similar designations by the IARC and other agencies, industry groups such as the American Chemistry Council successfully advocated for a review of the report by the National Academies in an effort to get the chemicals delisted and avoid further regulation. The outcome of the review is expected soon.

A QUESTION OF QUANTITY

Research is painting an increasingly complex picture of how the body reacts to environmental chemicals. In the past two years, debate has peaked about the behaviour of endocrine-disrupting chemicals. A 2012 review[5] of more than 800 studies concluded that it is "remarkably common" for these chemicals to induce biological responses at much lower doses than expected, and for the responses to be non-monotonic — that is, for higher doses not necessarily to produce greater effects than lower doses. The assertion has huge regulatory implications because safety testing for most chemicals is done not at the low doses at which agents occur in the environment, but at high doses. The results are extrapolated to lower doses, a methodology that would be unsound if non-monotonic behaviour were widespread.

These phenomena have been highly contentious. The US Environmental Protection Agency (EPA) has concluded that non-monotonic responses do occur but are uncommon. Its draft report is currently under review by a National Academies committee. Even so, the WHO, the United Nations Environment Programme and the US Endocrine Society all underscore the importance of low-dose and non-monotonic responses.

Another idea gathering steam is that people are particularly vulnerable at certain times, such as during gestation, puberty and pregnancy, and after giving birth. For example, a study[6] on human prostate stem cells implanted into mice indicated that early-life exposure to low doses of the endocrine-disrupting chemical bisphenol A, which is used in the manufacture of food containers and drink bottles, may increase the risk of a man developing prostate cancer.

Human studies, too, support 'windows of susceptibility'. The Child Health and Development Studies (CHDS) have found that women who have high levels of polychlorinated biphenyls (PCBs) in their blood immediately after giving birth have a tripled risk of developing breast cancer nearly two decades later[7]. The pesticide DDT has an even greater effect[8].

Another CHDS study[9] found that mothers of sons who developed testicular cancer in their thirties tended to have had a suite of DDT-related compounds in their blood when they gave birth. It is now becoming clear "that those really early exposures and events in life could influence the health trajectory of a lifetime", says epidemiologist and CHDS director Barbara Cohn.

"We've got to stop pouring this carcinogenic stuff out into the economy."

Dozens of environmental chemicals are regularly detected in people. Figuring out how these mixtures influence the risk of cancer and other problems may be the most puzzling question for researchers. Most studies investigate the effect of just one chemical at a time, and even that is tricky. "It is very unlikely for us to unravel what we need to know about chemicals until we can understand the implications of mixtures," says Cohn. "We're just learning how to do this."

Looking ahead, Cohn says that she plans to study whether chemical exposures reverberate into the third and fourth generations; the first great-grandchildren of the women who originally participated in the study will soon be born. Rodent studies have shown that exposure to toxic agents can increase the risk of cancer and other illnesses not only in the exposed animal and its offspring, but also in its offspring's offspring — often as a result of epigenetic alterations to gene expression[10].

IARC director Christopher Wild has coined the term 'exposome' to describe every exposure a person experiences during his or her lifetime — including chemicals, infectious agents, diet, social milieu and more. To better understand what might trigger cancer, Wild suggests that researchers hunt for the hallmarks of exposures, for instance in metabolic products or the pool of RNA molecules transcribed in certain cells.

A promising advance is that the genomes of some tumours have been found to contain distinct mutations when the person has been exposed to particular environmental risks such as tobacco or ultraviolet light. "The tumour starts to reveal its own secrets of its origins," Wild says. Wild has spent three decades researching cancer. During that time, he says, "I don't think anybody's ever asked me: 'Have we found out how to prevent it?' It's always, 'Have we found a cure?'"

Moreover, Wild and others say, prevention efforts usually focus on lifestyle changes for which the individual is responsible, such as stopping smoking and eating well, rather than regulatory changes that would place the responsibility on companies or governments to protect people from exposure to carcinogens.

A CALL TO PREVENT

The IARC's *World Cancer Report 2014* notes that cancer rates are rising fastest in developing countries (see page S64) and that some of the latest treatments will be too expensive for most people to access. It calls for a renewed commitment to prevention, including legislation limiting exposure to environmental carcinogens. "It just seemed an obvious conclusion that you couldn't treat your way out of cancer," Wild says.

Clapp agrees. He points to the promise of green chemistry, the development of safe molecules and techniques to replace harmful ones. An example is replacing dry-cleaning that uses perchloroethylene (classified by the IARC as probably carcinogenic) with soap-based 'wet-cleaning'. Clapp also applauds the European Union's chemical regulation, known as REACH, which puts the onus on companies to demonstrate a chemical's safety before it hits the market. "We've got to stop pouring this carcinogenic stuff out into the economy so that people don't get cancer in the first place," he says

China has started to do just that with air pollution. But curbs on exhaust and industrial emissions and fines for polluters have not yet had much effect: just 3 of the 74 cities the government monitors have met air-quality standards. For now, some citizens resort to wearing face masks outdoors and running air filters in their homes — but that only goes so far. "To clean up the air — that's the solution," Cohen says. "It's not to make people sit in their houses or walk around with masks on." ■

Rebecca Kessler *is a freelance science journalist in Providence, Rhode Island.*

1. Straif, K., Cohen, A. & Samet, J. (eds) *Air Pollution and Cancer* (IARC, 2013).
2. Doll, R. & Peto, R. J. *Natl Cancer Inst.* **66,** 1191–1308 (1981).
3. President's Cancer Panel *Reducing environmental cancer risk* (National Institutes of Health, 2010).
4. National Toxicology Program *Report on Carcinogens: Twelfth Edition* (HHS, 2011).
5. Vandenberg, L. N. *Endocrine Rev.* **33,** 378–455 (2012).
6. Prins, G. S. *et al. Endocrinology* **155,** 805–817 (2014).
7. Cohn, B. A., Terry, M. B., Plumb, M. & Cirillo, P. M. *Breast Cancer Res. Treat.* **136,** 267–275 (2012).
8. Cohn, B. A., Wolff, M. S., Cirillo, P. M. & Sholtz, R. I. *Environ. Health Persp.* **115,** 1406–1414 (2007).
9. Cohn, B. A., Cirillo, P. M. & Christianson, R. E. *Arch. Environ. Occup. Health* **65,** 127–134 (2010).
10. Guerrero-Bosagna, C. *et al. Reprod. Toxicol.* **34,** 694–707 (2012).

JACQUELINE KOCH/FRED HUTCHINSON CANCER RESEARCH CENTER

In Uganda, families dealing with cancer find support from fellow patients in treatment clinics.

DEVELOPING WORLD

Global warning

Much of the world is ill-equipped to cope with its rising cancer burden and are pushing prevention and screening.

BY ERIC BENDER

In the villages outside Bangalore in southern India "there's a lot of fear around cancer", says social epidemiologist Suneeta Krishnan. "Women know other women who have breast cancer, or died of cervical cancer. They have little awareness that early detection can lead to good outcomes — and a feeling that they'd rather just not know, because they couldn't afford treatment."

The concerns noted by Krishnan, who works with RTI International's Women's Global Health Imperative in San Francisco, California, are common in the developing world, where prevalence of cancer is climbing rapidly. Experts are raising the alarm over an incoming tidal wave of diagnoses in low- and middle-income countries (LMICs) that will be met with health-care resources that are starkly limited at best.

Of the 14 million people diagnosed with cancer worldwide in 2012, more than 60% live in Africa, Asia and Central and South America, according to the *World Cancer Report 2014*; these regions also account for about 70% of the world's 8 million cancer deaths[1].

Global cancer incidence is predicted to reach 25 million by 2032. The share of the burden borne by LMICs will almost certainly grow, say public-health experts, as populations expand, live longer and adopt the Western lifestyles that are associated with risks of numerous types of cancer (see 'Driving demographics'). "As life expectancies go up, cancer goes up, and these countries are totally unprepared to deal with it," says Mary Gospodarowicz, medical director of the Princess Margaret Cancer Centre in Toronto, Canada, and president of the Union for International Cancer Control (UICC) in Geneva, Switzerland.

Some nations have only a handful of oncologists, and doctors who leave to get trained in developed countries often stay abroad. Underlying these problems are severe limits on health-care funding. "Cancer care is a huge expenditure, and governments and private institutions universally are struggling to cover the costs," says Corey Casper, co-director of the Uganda Cancer Institute/Hutchinson Center Cancer Alliance in Seattle, Washington.

The dearth of health-care resources in developing countries is compounded by fears and stigmas associated with cancer. Public-awareness challenge number one is to dislodge the common conviction that cancer is an automatic death sentence.

In Latin American nations, women shy away from breast screening because they expect those diagnosed with cancer to die of the disease. "If they've only seen women die, they won't run forward to getting a mammogram," says Felicia Knaul, director of the Harvard Global Equity Initiative in Boston, Massachusetts.

Education campaigns must target not just patients or potential patients but also their families and the communities around them. One key is persuading men that the women in their families should get screened, because a wife, for example, may not feel free to take that step without her husband's permission. "Men can be obstacles," says Princess Dina Mired, director-general of the King Hussein Cancer Foundation in Amman. "We tell them that these are your daughters, your wives and your mothers, and you need to support them."

In addition to calling for ramped-up access to screening, advocates around the world are also speaking up on the need for improved end-of-life pain control. "There is zero access to morphine in the poorest countries," says Knaul. "The vast majority of cancer patients die in excruciating pain, and it isn't really an issue of money." Legal restrictions are often the cause; India is one of several countries that are thinking about easing constraints on the use of opiates in palliative care.

The successes of the global campaign against HIV/AIDS offer lessons. "Just as advocates acted up and spoke out on HIV, we need to act up and speak out on cancer and other chronic diseases," says Knaul.

PREVENTION IS A PRIORITY

One-third of the deaths from cancer in LMICS are preventable, according to the World Health Organization (WHO). Measures to reduce exposure to carcinogens such as tobacco; boost vaccinations against infections that can cause cancer; and encourage healthy lifestyles that lower cancer risks can help to avert the rising tide of cancer in LMICs, says Gospodarowicz. "And if you can prevent a lot of cancers, hopefully you will have more resources to deal with the cancers that are not preventable."

The WHO attributes just under one-quarter of global cancer deaths to tobacco, and the news on anti-smoking efforts is mixed. Encouragingly, 177 countries are implementing the WHO Framework Convention on Tobacco Control, says Hana Ross, head of international tobacco control research at the American Cancer Society in Atlanta, Georgia. The convention came into force in 2005 and offers a proven set of tools to reduce tobacco use, ranging from boosting health education and increasing taxation to establishing smoke-free zones and putting pictorial warnings on cigarette packages.

But putting the anti-smoking ideas into practice is another matter. "The battle is changing because of the strength of the tobacco companies, which can make it difficult to implement and enforce those provisions," Ross says. In China, which has about 350 million smokers (including more than half of all males) and where the government owns the world's largest tobacco firm, progress remains particularly difficult. Across Africa, as the population grows and people are increasingly targeted by marketing, "it's going to be catastrophic", she says.

Viral infections are another major contributor to cancer mortality. Up to 20% of cancer deaths in low-income countries arise from hepatitis viruses, which can cause liver cancer, or human papillomavirus (HPV), which is linked to cervical and other cancers. In recent progress against HPV, drug manufacturers are lowering the price of vaccines for low-income countries. Additionally, clinical trials have indicated that a single dose of HPV vaccine may offer the same level of protection against cervical cancer as the customary three doses[2], and is considerably easier to administer. Moreover, an unpublished phase III study showed that a Merck vaccine called V503 is effective against nine strains of HPV, with the potential to prevent about 90% of cervical cancers, as opposed to about 70% prevention from four strains covered by the company's current vaccine, Gardasil.

UNIQUE CHARACTER

Some aspects of cancer in the developing world contradict conventional understanding of the disease, and researchers do not yet know why. Breast cancer provides an example. "Women in the developing world often get breast cancer at a much younger age and in a much more aggressive form than women in the developed world," says Princess Dina. "These women marry early, have children early, breastfeed and probably don't use birth control. Those are supposed to be factors that prevent against breast cancer, so why are they getting this kind of cancer?"

Another major puzzle is why the number of cancers that are driven by infections is so high in the developing world. Some causes are well-known, such as a dearth of the pap-smear tests widely used in the West to identify HPV-linked abnormalities before they develop into full-blown cancers. But others remain a mystery.

In one effort, the Uganda Cancer Institute/Hutchinson Center Cancer Alliance is studying how newborn babies acquire the infections that can lead to cancer. The work has found that viruses such as Epstein–Barr (associated with Burkitt's lymphoma and other cancers) show up in almost all Ugandan children within the first two years of life and often can be detected for years afterwards, says Casper.

Cost-effective approaches to screening and diagnostics are also undergoing intense testing. For instance, a pilot programme in Uganda, run by the non-profit organization Imaging the World in Naalya, has successfully trained local health workers to use rugged portable ultrasound machines to help detect breast cancer, according to Casper. Suspicious lumps are given fine-needle biopsies, which can be analysed by equipment that is already widespread in Africa for use in HIV care.

Elsewhere, work is examining the effectiveness of screening that can be done with locally available resources — such as visual inspections of the cervix after dabbing it with acetic acid. The approach is not as accurate as pap smears, but is much less expensive and demanding of expertise, and it means that abnormal cells can be immediately removed by cell-freezing methods. A randomized study[3] of 150,000 Indian women indicated that the technique can cut cervical-cancer deaths by 31%.

Even if health care is available and affordable, front-line health-care workers in LMICs are often not properly trained to diagnose and treat cancer. And because of cultural norms, women in many countries may not permit a male doctor to examine them. As a result, the doctor may have no option but to make a diagnosis based on a woman's description of her symptoms. This restriction is one of the reasons that a very high proportion of women arriving at major care centres already have advanced tumours that have never been examined by a health-care worker.

'Twinning' partnerships, in which health-care institutions in developed nations offer expertise to colleagues in developing countries, offer a way to improve cancer care. The concept could also be applied within a country, with national and state centres working with smaller, local centres, Knaul suggests. Given proper training and supervision, it may be possible to complete certain drug treatments in the local centres, making cancer therapy practical for patients who live far from clinics. And the near ubiquity of mobile phones is making telemedicine more practical. "We have a mechanism to make sure patients don't get lost in follow-up," says Edward Trimble, director of the US National Cancer Institute's Center for Global Health in Rockville, Maryland.

Trimble also points to promising early experiments using mobile phones to take images of tumours at remote locations and send them to oncology specialists. This approach could aid in diagnosing the eye cancer retinoblastoma, for example, and apps are being developed to help diagnose other types of cancer as well. "There aren't any apps that have been shown to be 100% effective around the world, but this is an important area of research," says Trimble.

SETTING GLOBAL GOALS

The worldwide conversation about cancer is changing, says Tezer Kutluk, a paediatric oncologist at Hacettepe University's Faculty of Medicine in Ankara. Since it became the leading killer worldwide in 2011, "cancer is more and more on the agenda of LMICs, and its priority is increasing", he says.

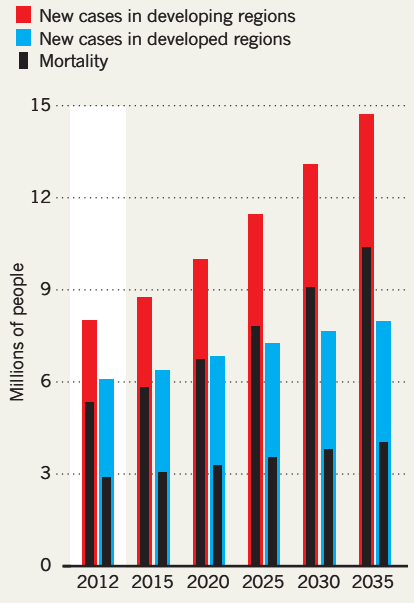
On a positive note, he adds, the WHO, governments, public and private institutions, and other stakeholders are learning to work together more efficiently on both cancer and other non-communicable diseases — one example being the UICC's global public-awareness campaign held each year on World Cancer Day.

Given predictions of an incoming wave of cancer cases in LMICs and in view of the constraints on resources, "there's a lot of controversy and a lot of effort right now to find the best buys for LMICs to invest in", says Gospodarowicz. "You need to have all the pieces in the puzzle for cancer control."

The quest to solve this puzzle, she says, is accelerating the demand for evidence of the most effective and efficient ways to control cancer — and the demand to share the results of that science globally. "The world is now thinking not just about developing new cancer drugs or X-ray machines or other technologies but about how to actually apply them," she says. "That's a very exciting development." ■

DRIVING DEMOGRAPHICS

The number of people who will develop and die from cancer is predicted to climb more steeply in developing countries than in developed nations as populations grow and age. Lifestyle changes, such as Westernization, are likely to boost the numbers further.



SOURCE: WHO

Eric Bender *is a freelance science writer based in Newton, Massachusetts.*

1. Stewart, B. W. & Wild, C. P. (eds) *World Cancer Report 2014* (IARC, 2014).
2. Safaeian, M. *et al. Cancer Prev. Res.* **6,** 1242–1250 (2013).
3. Shastri, S. S. *et al. J. Natl Cancer Inst.* **106,** dju009 (2014).

BRENDAN MONROE

BIOINFORMATICS

Big data versus the big C

The torrents of data flowing out of cancer research and treatment are yielding fresh insight into the disease.

BY NEIL SAVAGE

In 2013, geneticist Stephen Elledge answered a question that had puzzled cancer researchers for nearly 100 years. In 1914, German biologist Theodor Boveri suggested that the abnormal number of chromosomes — called aneuploidy — seen in cancers might drive the growth of tumours. For most of the next century, researchers made little progress on the matter. They knew that cancers often have extra or missing chromosomes or pieces of chromosomes, but they did not know whether this was important or simply a by-product of tumour growth — and they had no way of finding out.

"People had ignored it for a long time, primarily because it's really hard to understand," says Elledge, of Brigham and Women's Hospital in Boston, Massachusetts. "What we didn't know before is that it's actually driving cancer."

Elledge found that where aneuploidy had resulted in missing tumour-suppressor genes, or extra copies of the oncogenes that promote cancer, tumours grow more aggressively (T. Davoli *et al. Cell* **155,** 948–962; 2013). His insight — that aneuploidy is not merely an odd feature of tumours, but an engine of their growth — came from mining voluminous amounts of cellular data. And, says Elledge, it shows how the ability of computers to sift through ever-growing troves of information can help us to deepen our understanding of cancer and open the door to discoveries.

Modern cancer care has the potential to generate huge amounts of data. When a patient is diagnosed, the tumour's genome might be sequenced to see if it is likely to respond to a particular drug. The sequencing might be repeated as treatment progresses to detect changes. The patient might have his or her normal tissue sequenced as well, a practice that is likely to grow as costs come down. The doctor will record the patient's test results and medical history, including dietary and smoking habits, in an electronic health record. The patient may also have computed tomography (CT) and magnetic resonance imaging (MRI) scans to determine the stage of the disease. Multiply all that by the nearly 1.7 million people diagnosed with cancer in 2013 in the United States alone and it becomes clear that oncology is going to generate even more data than it does now. Computers can mine the data for patterns that may advance the understanding of cancer biology and suggest targets for therapy.

Elledge's discovery was the result of a computational method that he and his colleagues developed called the Tumor Suppressor and Oncogene Explorer. They used it to mine large data sets, including the Cancer Genome Atlas, maintained by the US National Cancer Institute, based in Bethesda, Maryland, and the Catalogue of Somatic Mutations in Cancer, run by the Wellcome Trust Sanger Institute in Hinxton, UK. The databases contained roughly 1.2 million mutations from 8,207 tissue samples of more than 20 types of tumour.

The researchers selected a set of parameters that helped to identify the genes they were looking for, such as the mutation rate or the ratio of benign mutations to those that cause a gene to stop functioning. They then applied statistical classification methods to differentiate between suppressor genes and oncogenes. About 70 suppressor genes and 50 oncogenes were already known for these tumour types, but Elledge and his colleagues increased that to about 320 and 200, respectively (although

➲ NATURE.COM
Find out more about software design for bioinformatics:
go.nature.com/auca5y

that number could fall, because some genes could turn out to be false positives). They also identified pathways in the growth process that might make good drug targets.

Making this sort of finding requires large data sets. "Any individual cancer cell's a mess, but if you look at enough tumours, you get a pattern," Elledge says. "The only way you can figure this out is if you look at them globally."

EASY TO USE

Analysing the genomes of 8,200 tumours is just a start. Researchers are "trying to figure out how we can bring together and analyse, over the next few years, a million genomes", says Robert Grossman, who directs the Initiative in Data Intensive Science at the University of Chicago in Illinois. This is an immense undertaking; the combined cancer genome and normal genome from a single patient constitutes about 1 terabyte (10^{12} bytes) of data, so a million genomes would generate an exabyte (10^{18} bytes). Storing and analysing this much data could cost US$100 million a year, Grossman says.

To make it easier to access whatever subset of data researchers need, Grossman and his colleagues have developed Bionimbus, a cloud-based, open-source platform for sharing and analysing genomic data from the Cancer Genome Atlas.

The results can be powerful. Megan McNerney, a pathologist at the University of Chicago, used Bionimbus to track down a gene involved in acute myeloid leukaemia (AML). Scientists already knew that some patients with the disease had lost part of chromosome 7, but could narrow down the gene involved only to 15–20 candidates. McNerney selected 23 patients from the database and used the computer to compare their RNA sequences to see if something might be missing. She discovered that one copy of the gene *CUX1*, which normally encodes a tumour-suppressor protein, had been deleted in these patients (M. E. McNerney *et al. Blood* **121,** 975–983; 2012). Testing in fruit flies and mice showed that removal of one copy of the gene led to an overgrowth of certain blood cells and, eventually, to leukaemia. Her discovery may not have produced a cure for AML, but it has increased the understanding of a disease for which the median survival time has been stuck at less than a year for four decades, and it might also lead to more-accurate prognoses.

McNerney says that even her small-scale project has shown the benefits of mining data. "It's transforming cancer biology enormously," she says. "Big data has made leaps that we couldn't make otherwise."

Genomics — and data from other '-omics, such as proteomics and epigenomics — are not the only sources of data being sifted. The American Society of Clinical Oncology (ASCO) in Alexandria, Virginia, is developing a platform called CancerLinQ, which trawls through patients' electronic health records. These records increasingly include genomic data, as well as diagnoses and notes on treatment, and measures of how well patients are responding to therapy. The system has gathered records from 177,000 people with breast cancer for a pilot project. Developers hope that the system will be fully operational by the summer of 2015, with other solid tumours to follow.

Clifford Hudis, a breast-cancer specialist at the Memorial Sloan Kettering Cancer Center in New York and president of ASCO, says that CancerLinQ could make discoveries missed by clinical trials. As approved drugs are deployed more widely, the system could gather data on side effects, drug interactions and outcomes in different patient populations. It might also notice, for instance, if doctors stray from US Food and Drug Administration guidelines for drug dosage, based on their assessment of how the dose affects their patients. "If there are 100 cases in a row of doctors independently disregarding the guideline, it helps to teach the computer that the guideline's wrong," Hudis says. The computer might discover, for instance, that doctors get better results when they adjust the dosage according to the patient's age.

> ***"There are some fundamental challenges being caused by our ability to capture so much data."***

Discoveries can also be made from combining genomics and standard medical-imaging records. "High-performance computing and big data are enabling us to look across modalities," says David Foran, a pathologist and head of informatics at the Rutgers Cancer Institute of New Jersey in New Brunswick. The centre produces high-resolution digital images of tissue samples and compares them between patients, looking for patterns that might aid prognosis. It expects to generate 40,000–100,000 images.

Researchers might see genetic clues indicating that some patients will respond to a particular drug therapy, for instance, and then look at their CT and MRI scans to see whether changes in the cancer match up with the genetic prediction. Or they might find a correlation between mutations, therapy choice and smoking history. "The computer program can simultaneously look at the patterns in all of them," Foran says.

Comparing so much data greatly expands doctors' expertise, Foran adds. "When you go to see a physician, especially an oncologist, you're relying on his past experience. What we're doing now is training the computer to look at large cohorts of thousands and hundreds of thousands." It is as if the doctor were making treatment decisions based on personal experience of hundreds of thousands of patients.

Gene sequences and electronic health records are new sources of data, but there is a lot of historical information available, too. Johns Hopkins Hospital in Baltimore, Maryland, for instance, has paper-based pathology reports that date back to its opening in 1889. Before it switched to computer records in 1984, the hospital generated more than half-a-million records. Every US state has years or decades of historical cancer records, as do other countries. Denmark, for instance, has cancer records going back to 1943. And Public Health England last year launched a database of all cancers currently being diagnosed across the country, including 11 million records going back 30 years. Adding all that history into the mix widens the field of possible clues that computers can search through.

HARD TO ANALYSE

But it is the new technologies that are creating an information boom. "We can collect data faster than we can physically do anything with them," says Manish Parashar, a computer scientist and head of the Rutgers Discovery Informatics Institute in Piscataway, New Jersey, who collaborates with Foran to find ways of handling the information. "There are some fundamental challenges being caused by our ability to capture so much data," he says.

A major problem with data sets at the terabyte-and-beyond level is figuring out how to manipulate all the data. A single high-resolution medical image can take up tens of gigabytes, and a researcher might want the computer to compare tens of thousands of such images. Breaking down just one image in the Rutgers project into sets of pixels that the computer can identify takes about 15 minutes, and moving that much information from where it is stored to where it can be processed is difficult. "Already we have people walking around with disk drives because you can't effectively use the network," Parashar says.

Informatics researchers are developing algorithms to split data into smaller packets for parallel processing on separate processors, and to compress files without omitting any relevant information. And they are relying on advances in computer science to speed up processing and communications in general.

Foran emphasizes that the understanding and treatment of cancer has undergone a dramatic shift as oncology has moved from one-size-fits-all attacks on tumours towards personalized medicine. But cancers are complex diseases controlled by many genes and other factors. "It's not as if you're going to solve cancer," he says. But big data can provide new, better-targeted ways of grappling with the disease. "You're going to come up with probably a whole new set of blueprints for how to treat patients." ■

Neil Savage *is a freelance science and technology writer based in Lowell, Massachusetts.*

PERSPECTIVE

Learning to share

Genomics can provide powerful tools against cancer — but only once clinical information can be made broadly available, says **John Quackenbush**.

The unveiling of the draft sequence of the human genome in 2000 was met with enthusiastic predictions about how genomics would dramatically change the treatment of diseases such as cancer. The years since have brought a 100,000-fold drop in the cost of sequencing a human genome (to just a few thousand US dollars), and the time needed to sequence it has been cut from months to little more than a day. Researchers can therefore now generate unprecedented quantities of data to help in the battle against cancer (see page S66).

So far, however, our expanded data-generation capacity has not transformed medicine or our understanding of the disease to the degree that some expected. A major contributor to this disappointing outcome has been the failure to deal effectively with the problem of capturing and sharing appropriate clinical data on large collections of samples.

The ultimate goal of cancer researchers is to deliver actionable point-of-care information to doctors treating patients. This means, for example, producing easy-to-read reports that detail the associations between a patient's disease state and their probable response to available therapeutics — associations that are defined by a variety of clinical and genomic attributes and that should be supported by a large, well-curated knowledge base. This information can then help doctors to make rapid decisions about which course of therapy is likely to work best for each patient.

Research has already established associations between a few gene variants or gene-expression profiles and clinical endpoints such as drug response. But given the ability to generate large-scale genomic profiling data, they have identified many fewer variants than might have been expected. This shortfall can be attributed to failings in current clinical-research paradigms.

The fundamental design of most clinical and translational-research studies involves comparisons between well-defined patient cohorts. Researchers may divide patients into groups on the basis of outcome — for example, response to a therapy — and ask whether there are genomic features such as mutations or patterns of gene expression that can robustly distinguish between responders and non-responders. Or they can define patient groups according to genomic status and then ask whether there are meaningful differences in some relevant endpoint, such as survival. Cancer research has produced thousands of such genomic studies, with data on hundreds of thousands of patients. But very few of the published studies have been thoroughly validated and fewer still have proved clinically useful.

Although researchers have rushed to generate genomic data, that alone is not sufficient to advance the field. One challenge is to develop analytical methods that are effective for huge amounts of genomic data. In particular, better methods are needed to 'normalize' the data generated by different technologies or at different sites, so that results can be compared across studies — a problem that may seem trivial but that nevertheless has defied a general solution. Methods are also needed to synthesize different types of information more effectively to make predictions, including ways to model the complex interacting networks of factors that drive disease. And standards must be developed to support reproducible research, facilitating validation of the results of any single study in the context of a collective body of data.

But the greatest barrier to the use of 'big data' in biomedical research is not one of methodology. It is, rather, the lack of uniform, anonymized clinical data about the patients whose samples are being analysed. Without such data, even defining experimental cohorts is difficult, and there is a risk of missing potentially obvious confounding factors. Unfortunately, nearly every published study lacks the clinical data to address fundamental research questions fully or to allow the findings of one study to be validated in others.

The first step towards solving this problem is to develop more flexible patient-consent procedures so as to allow the broad use of anonymized clinical data in research. This is particularly important because, at the start of a study, researchers may not know which variables could be important for defining a relevant cohort or could turn out to be confounding an analysis.

The second step is to develop hospital and laboratory computational-infrastructure and data-security protocols to improve the sharing, access and fair use of clinical data. A major barrier to reproducing results is that publicly available data sets rarely include the right clinical information to define appropriate cohorts or test the relevance of a genomic signature.

PUBLICLY AVAILABLE DATA SETS **RARELY INCLUDE** THE RIGHT CLINICAL INFORMATION TO **DEFINE** APPROPRIATE COHORTS OR TEST THE **RELEVANCE** OF A GENOMIC SIGNATURE.

And, finally, the culture of data sharing must change. Although the publication of the results of genomics studies generally requires the sharing of genomic data, the sharing of clinical data is frequently limited to a bare minimum: just the details described in a manuscript. Even common clinical variables, such as patients' sex, treatment history, smoking history, ethnicity or even standard disease subtype are often not provided. The absence of such key information again makes it difficult to reproduce the results of an analysis or to validate other published data sets.

Big data has tremendous potential to provide fresh insights into diseases such as cancer. But that potential will be realized only by tackling how best to share the clinical information necessary to interpret it. And developing a more complete understanding is essential if we are ultimately to create the knowledge base necessary to provide clear, concise, reliable and actionable information to doctors and their patients. ■

John Quackenbush *is director of the Center for Cancer Computational Biology at the Dana-Farber Cancer Institute in Boston, Massachusetts.*
e-mail: johnq@jimmy.harvard.edu

BRENDAN MONROE

BIOLOGY

Three known unknowns

Even as cancer therapies improve, basic questions about drug resistance, tumour spread and the role of normal tissue remain unanswered.

BY KATHERINE BOURZAC

In 1996, Charles Sawyers designed early clinical trials for one of the first drugs aimed at a cancer-specific genetic mutation. The drug was imatinib, the cancer was chronic myeloid leukaemia and Sawyers — a clinical oncologist at the Memorial Sloan Kettering Cancer Center in New York — saw patients who had been debilitated by the disease rapidly improve when given the medicine. "It was unbelievably satisfying," he says.

Unfortunately, he then saw many of those cancers come roaring back as they became resistant to the drug.

The experience with imatinib has given cancer biologists mixed messages. The medicine, now marketed by Novartis in Basel, Switzerland, as Gleevec or Glivec, highlights the potential of personalized medicine. Figuring out what mutation caused the disease and designing a drug to target it was a technological triumph, and it was followed by two further drugs to combat the emerging drug resistance.

But treating cancer by chasing mutation after mutation with drug after expensive drug is not a sustainable model — not least because few cancers other than leukaemia have simple, known genetic causes. "When we know the mutations and can get to a treatment strategy it's exciting," says Sawyers. But so far in the age of gene sequencing, he adds, "we've grabbed the low-hanging fruit".

Biologists now know a huge amount about cancer — much more than they did even ten years ago. About 500 genes have been implicated in the disease, and the list is growing. There are also about 100 approved cancer drugs, some of which, like imatinib, specifically target mutations in those genes, on top of older therapies such as surgery and radiation.

But all this knowledge is not enough: even in countries where people have access to the newest therapies, improvements in death rates have slowed. Up to half of cancers could be prevented by changes in diet and exercise, encouraging people to stop smoking and eliminating environmental risks such as pollution, but other gains will be harder. To conquer cancer, researchers will need to answer some basic scientific questions. Here, *Nature* looks at three of the most pressing.

HOW CAN DRUG RESISTANCE BE OVERCOME?

To combat resistance, researchers are studying the cancer genome, coming up with new ways to design drugs, concocting

combination therapies — and even looking back to Darwin's theory of evolution.

"Seen through a Darwinian lens, the tumour is an ecosystem, a mixture of cells that are continuously mutating," says Paul Workman, head of cancer therapeutics at the Institute of Cancer Research in London. "You put into that mix a very strong selective pressure, which is the drug." At that point it becomes survival of the fittest. Many cells die; others use a combination of strategies to survive and thrive. These may include producing protein pumps that flush the drug out, increasing the rate of DNA repair or using an alternative molecular pathway to restore whatever function the drug blocks. Targeted drugs contribute to the genetic complexity: "These therapies themselves may be driving tumours to become more heterogeneous," says Charles Swanton, a medical oncologist at Cancer Research UK's London Research Institute.

A better understanding of the underlying genetic diversity of tumour cells may help researchers to work out how to tackle drug resistance. Swanton and others are therefore exploiting ever-faster and cheaper DNA-sequencing technologies. So far, Swanton says, it looks as though every tumour has a set of core mutations that are shared by all its cells. He calls these the tumour's 'trunk'. Subpopulations of cells within the tumour have their own unique sets of shared mutations; he calls these subpopulations 'branches'. Therapy prunes some branches while sparing others, which then repopulate the tumour.

Researchers are now trying to look at tumour evolution in patients. One study, called TRACERx (Tracking Cancer Evolution through Therapy), will allow Swanton and a large group of collaborators to observe 850 people with lung cancer from diagnosis through therapy. Biopsies are taken from multiple spots within tumours both before and after treatment, then analysed by sequencing the parts of the tumour genomes that code for proteins. Comparing these biopsies should identify which mutations are associated with drug resistance. These kinds of studies may help geneticists to write what Swanton calls "an evolutionary rulebook of cancer" that can be used to predict tumour evolution without having to do repeated sequencing studies to get future patients on the right therapies.

Other researchers caution that genetics will provide only part of the picture of tumour heterogeneity and drug resistance. Variations in how tumours use these genes — the way they are regulated and expressed — also enable tumours to develop drug resistance. "Cells that are not intrinsically resistant to a drug will rewire their circuitry during treatment to become resistant" without any genetic changes at all, says cell biologist Joan Brugge at Harvard Medical School in Boston, Massachusetts.

Even without a full understanding of the way that tumours evolve in the face of chemotherapy, researchers are coming up with ways to overcome resistance. Using a combination of drugs can reduce a tumour's options. Here, scientists take inspiration from the success of the antiretroviral cocktails that keep HIV in check. Like cancer, HIV has tremendous genetic diversity and evolves rapidly, but the right cocktail of drugs has transformed HIV infection from a death sentence for many into a manageable, long-term condition. Cancer presents a tougher challenge. HIV has just nine genes, compared with our approximately 20,000, making human cancer cells much more complex. Researchers are still trying to figure out how to make smart combination therapies that really work.

"Cells that are not intrinsically resistant to a drug will rewire their circuitry during treatment to become resistant."

James Doroshow, head of cancer treatment and diagnosis at the US National Cancer Institute (NCI) in Bethesda, Maryland, believes that the best way to figure out combination therapies is to test the possibilities through brute force. The NCI has been testing 5,000 drug combinations against 60 cancer cell lines *in vitro*; promising candidates are then screened for toxicity in mice. The results have not yet been published, but Doroshow says that new and unexpected combinations are showing up.

Workman's group is using computer models of gene networks to sort through thousands of possible drug combinations and genes to find likely synergies. He agrees that combination therapy is the only way to overcome resistance, but thinks that new drugs are also needed. He estimates that just 5% of known cancer genes are targeted by drugs. "If we can't make drugs against the other 95%," he asks, "how on Earth are we going to build the combination therapies that will lead to a cure?"

To make matters more difficult, some cancer-causing mutations work by silencing the tumour-suppressor genes that normally help to stop tumours from forming. Developing a drug to block the absence of something is a major challenge, says Workman. And some of the genes associated with cancer make proteins whose structures are unknown; without the structure, chemists have nothing to go on. Many cancer genes therefore remain, for the time being at least, untargetable.

HOW ARE HEALTHY TISSUE AND GENES INVOLVED?

Cancer is caused not just by bad cells or bad genes, but also by good ones not doing the right thing — an aspect of cancer that is highly complicated to study and to combat.

Mina Bissell, a bioengineer at the Lawrence Berkeley National Laboratory in California, sees potential in an area that has been largely ignored by the pharmaceutical industry: the supportive structures and non-cancer cells in and around tumours, called the tumour microenvironment. Signals from this microenvironment can stop a cell that has cancer-causing mutations from becoming cancerous, and putting a tumour cell in a different environment can render it benign. "We need to therapeutically fix the microenvironment," she says.

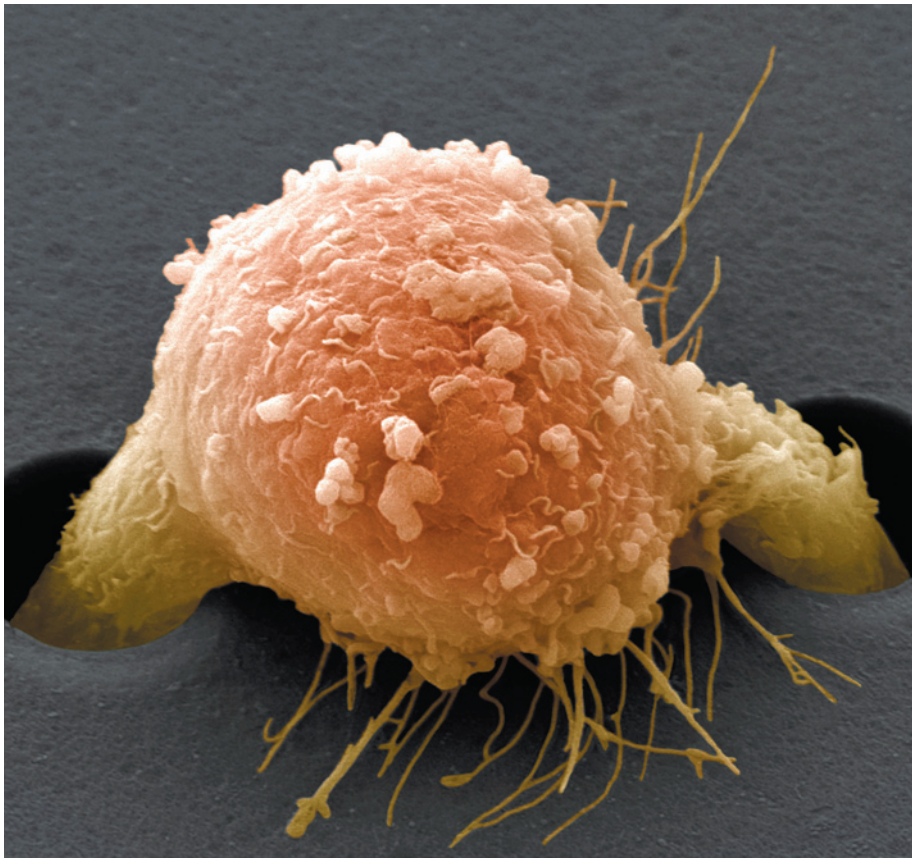
Jacqueline Lees, associate director of the Koch Institute for Integrative Cancer Research at the Massachusetts Institute of Technology in Cambridge, agrees. It is important, she says, to think not just about killing cancer cells but also about targeting the processes that support them, the interactions between tumour and non-tumour cells and the immune system. Tumours cannot thrive without certain kinds of signalling patterns from their neighbouring cells. "Traditional drug screening has missed that," Lees says.

Lees is studying a process called cancer dormancy. Some cancer cells are quiescent during the tumour's boom times but can become reactivated if left behind after surgery. Lees is trying to determine how this works, and what role normal cells have in the dormancy and activation process.

Quiescent tumour cells are less vulnerable to chemotherapy because the treatments are aimed at dividing cells, and quiescent ones are not taking part in the normal cell cycle of growth and division.

What signals from the microenvironment are needed to wake these cells? Lees says that inflammation — a sign of activation of the immune system — after surgery and treatment may jump-start quiescent cells. Then, a tumour cell has to warm up its engines. "It has to switch back on a whole protein program that takes 48 hours, and during that time it's taking advantage of interactions with normal surrounding cells," she says. The right drugs in the right order could help to control this process. Chemotherapy could be preceded by drugs that wake up the quiescent cells. Then drugs that suppress crosstalk could prevent them from restarting a tumour after therapy.

If it can be revved up, the immune system might be able to do this itself. Cancer cells put up many defences against attack from the immune system, including expressing receptors that deactivate an important class of immune cell. In the past few years, drugs that block these receptors have generated much excitement in clinical trials. These drugs, called immune checkpoint inhibitors, seem to let loose the immune system's natural cancer-fighting activity. But they do not work in all patients, and researchers have yet to figure out why, says Sawyers. For these



STEVE GSCHMEISSNER/SPL

Breast-cancer cells can become motile and start to spread around the body.

patients, a combination of drugs that target the tumour, the support cells and the immune system "could save the day", he says.

Lees agrees. Successfully treating cancer may require attacking tumours on multiple fronts, she says, with conventional therapies bolstered by new ones that activate the immune system and silence certain interactions between tumours and their environment.

HOW DOES CANCER SPREAD?

The cause of most cancer deaths — about 90% — is not the primary tumour, but secondary tumours called metastases that have developed elsewhere in the body. Sometimes these metastatic tumours become apparent decades after a patient was thought to be cured. So a better understanding of metastasis would help to prevent a great deal of cancer deaths. "We need to focus more on secondary tumours," says Ann Chambers, director of translational breast-cancer research at the London Health Sciences Centre in Canada.

The beginning of the metastatic process is by now pretty well understood, says Robert Weinberg, a cancer biologist at Whitehead Institute for Biomedical Research in Cambridge, Massachusetts. Some cancer cells become motile and aggressive, and enter the bloodstream. Some exit the circulation at distant sites, and a fraction of these can start new tumours there (C. L. Chaffer and R. A. Weinberg *Science* **331,** 1559–1564; 2011). What is incredible, says Weinberg, is that any of these cells live, let alone seed new tumours. "The big remaining mystery is how cancer cells are able to adapt and make a living in a distant tissue," he says. The environment in the brain or bone marrow, say, is very different from that in the prostate or breast, where the tumour may have started its journey. There may be different levels of glucose and oxygen, or the foreign tissue might be more or less acidic. Tumour cells are thought to be dependent on growth factors, protein signals and other encouraging messages in their native tissue. Still, some cells form new tumours at other sites. Weinberg speculates that this involves not mutations, but extensive changes in gene expression.

➲ NATURE.COM
To read about the insights from recent research, see:
go.nature.com/lsysrq

Perhaps even more puzzling than how cells can thrive in a new place is what they do in the time between their arrival and the growth spurt that initiates secondary disease. "Cancer cells make it to a distant organ, escape the bloodstream — and then they sit there for ten years while nothing happens," says tumour biologist Klaus Pantel of the University Medical Center Hamburg-Eppendorf in Germany. "Something keeps them from proliferating, and then something activates these cells."

What those signals are is unknown. And not enough people are studying the problem, says Chambers. "Tumour dormancy is frustrating — like watching paint dry," she says. "People don't want to watch tumours not grow."

There are other barriers to studying metastasis. Secondary tumours that have grown large enough to cause health problems and be detected are often not biopsied because patients are in fragile health. And it is difficult to get a picture of the early stages of growth of metastases: they are too small to show up in imaging scans. Furthermore, Chambers notes, even when people have ideas for drugs to prevent or slow tumour spread, today's clinical trials are not designed to show this effect. Trials tend to enrol patients with advanced disease and established metastatic tumours. The potential of a drug to prevent the spread of cancer cannot be seen in this group because it is too late, she says.

To treat cancers that have already spread will require knowing more about the mechanics of metastasis. First, researchers must figure out which of the heterogeneous mix of cells in a tumour are capable of spreading, and how they differ from the other cells.

Brugge has developed a way to find these tumour-initiating cells. Researchers in her lab take about 100 cells from a primary tumour biopsy, separate them, clone them and give each its own genetic bar code. They then maintain one set *in vitro* and inject another into mice, says Brugge. Cells with metastatic tendencies will grow into tumours, and once they do, the researchers remove them and note, using the bar codes, which cells started the new tumours. Brugge can then go back to the cells in culture to study what differentiates them from cells that did not metastasize.

Another approach involves looking for tumour cells in patients' blood, cancer's metastatic thoroughfare. These circulating tumour cells may hold some answers to the mystery of metastasis — their ranks must include some that will form secondary tumours. Once isolated, they can be sequenced and imaged, and the expression of their genes can be compared with that of the primary tumour's.

Many key details of cancer biology remain elusive, but new technologies are helping researchers to gain access to them. Quickly advancing genomic and bioinformatics techniques are helping to overcome drug resistance by predicting which drugs to use and in what combinations; new models are providing insight into the interactions between normal and cancerous tissues; and metastatic cells can now be found before they make tumours.

Given this progress, veteran researchers find reason for optimism. "Twenty years ago, I would have thought some of these problems were intractable," says Doroshow. "But now, I don't." ■

Katherine Bourzac *is a freelance science writer in San Francisco, California.*

TURNING POINT Fellowship lets physicist explore exotic states of matter **p.653**

FACEBOOK The latest in careers advice and information **www.facebook.com/naturejobs**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

VISUAL MOZART/GETTY

MEDICAL RESEARCH

Gene-therapy reboot

Safer, more efficient methods to deliver therapeutic genetic material are generating jobs for savvy scientists.

BY LAURA CASSIDAY

In the early 2000s, gene therapy seemed to be on life support. The once-promising technique, which uses engineered viruses and other methods to shuttle genes into human cells to fix DNA errors, faced a setback after an 18-year-old man died during a clinical trial in 1999. Later analysis showed that the virus carrying the DNA fixes had triggered a massive immune reaction that caused the man's organs to shut down.

But after years of troubled times, the field is on the move again. Success stories, an infusion of venture capital and regulatory approval in the European Union have revitalized the field and made scientists experienced in the technology hot commodities once more. "The field has really started to take off again," says Harry Gruber, chief executive of Tocagen, a gene-therapy company based in San Diego, California. The number of people working in gene therapy has grown enormously, he says, and that is going to continue.

Gene therapies generally use viruses to deliver the gene of interest to cells. The viruses — retroviruses, lentiviruses and adeno-associated viruses, among others — are defanged by removing harmful genetic sequences then engineered to contain the therapeutic gene sequence. When the virus encounters human cells, it inserts its genetic material into the human genome.

The first gene-therapy recipient, in 1990, was a four-year-old girl who had a genetic disorder known as adenosine deaminase (ADA) deficiency. The treatment, which used a retrovirus, seemed to work, although the effects were temporary and the girl continued to need treatment.

Still, the trial established that the procedure was safe and at least moderately effective, and the following decade saw the technique burgeon in hundreds of trials. But then the 18-year-old man died, and in 2002, a trial in France was shut down when a boy with an immune disorder known as X-linked severe combined immunodeficiency (X-SCID) developed leukaemia because of how the retrovirus had inserted itself into the genome. Three other boys in the French trial, and one in a similar UK SCID trial, ultimately developed leukaemia. Regulatory agencies in France, Germany, the United States and the United Kingdom suspended all trials that used techniques similar to that in the French study.

The X-SCID trial did have some successes, however. Nine of the ten boys in the French trial who received the therapy were cured of their deadly disease. And of the four who developed leukaemia, three were successfully treated with chemotherapy.

After a careful risk–benefit analysis, the agencies eventually lifted the restrictions and issued new safety guidelines. But investors and the public were badly shaken. Concerns about playing God resurfaced. Many first-generation gene-therapy companies folded, because the liabilities of the technique seemed to overwhelm the returns.

Yet researchers persevered. They worked to develop vectors that would insert themselves into the genome more precisely, and thereby avoid disrupting important genes. Gradually, success stories started to emerge.

At the University of Pennsylvania in ▶

Philadelphia, for example, researchers led by Carl June, director of translational research at the Abramson Family Cancer Research Institute, have seen promising results in clinical trials for chronic lymphocytic leukaemia (CLL) and acute lymphoblastic leukaemia (ALL).

Their approach involves isolating a patient's T cells and infecting them with a lentivirus that causes the cells to produce an antibody-like protein, or chimeric antigen receptor (CAR), on their surface, which binds to a protein on the surface of B cells. When the modified T cells were infused back into patients' bodies, they multiplied and attacked cancerous B cells. Almost half of the 32 adults with CLL responded to the therapy, and 19 of 22 children and all five adults with ALL experienced complete remissions, although some have since relapsed.

INDUSTRY INTEREST

In 2012, the university and the pharmaceutical firm Novartis in Basel, Switzerland, announced a partnership to study and commercialize the technology. The partnership, which will grant Novartis an exclusive worldwide licence to therapies based on CAR technology, will help to build a Center for Advanced Cellular Therapies on the university's campus. It will also fund the establishment of new labs and the hiring of scientists at the bachelor's, master's and PhD levels, says Bruce Levine, director of the university's clinical cell and vaccine production facility. He says that the alliance is an example of a growing trend of partnerships between industry and academia that are likely to increase demand for researchers. Another partnership that garnered media attention was the 2013 collaboration between the biotechnology company Celgene in Summit, New Jersey, with the Baylor College of Medicine in Houston, Texas, to develop gene therapies for cancer.

"We are ramping up at an extraordinary rate."
Bruce Levine

"We are ramping up at an extraordinary rate, due in large part to the Novartis alliance, but also to our other cell- and gene-therapy programmes," says Levine. "I half-joke to my colleagues when we're going to meetings that I should wear a badge that says, 'Ask me about the positions we have open.'" He says that "dozens" of positions are open, and that they are looking for people with a degree in a biology-related field and experience in developing and manufacturing gene and cell therapies. Also, strong record-keeping skills are a must owing to the extensive documentation required for the manufacturing of gene-therapy products for clinical trials.

People who want to move into gene-therapy research would do well to consider eye disorders, says Robert MacLaren, a surgeon at the University of Oxford's Nuffield Laboratory of Ophthalmology in Britain. "About one-third of genetic diseases manifest themselves in the retina," he says. Retinal cells are also easy to access and can be efficiently targeted with adeno-associated viruses (AAV), which are less likely to trigger an immune reaction or cause cancer than other viruses. "We're in the early stages, but undoubtedly the eye will be the most logical target organ for gene therapy," he says. "And I think what we learn about gene therapy in the eye will help us to apply the techniques much more effectively to other diseases."

MacLaren's trial of gene therapy for choroideremia, a rare, inherited form of blindness, triggered much excitement when the results were published in *The Lancet* in March (R. E. MacLaren *et al. Lancet* **383,** 1129–1137; 2014). Choroideremia is caused by a mutation in a gene called CHM that causes pigment cells in the retina to gradually stop working and die. MacLaren's team inserted a functional copy of CHM into an AAV and then injected it into one retina in each of six men with varying degrees of visual acuity. Five of the men showed improvements in their ability to see a dim light in the dark, and the two patients with the most severe choroideremia were able to read additional lines on an eye chart. These improvements have been sustained in the two years since the single treatment.

Researcher Samantha de Silva recognized the enormous potential of gene therapy for the eye. She had earned a medical degree and was completing her residency in ophthalmology at Oxford Eye Hospital, UK, when her career path veered in an unexpected direction. She heard about MacLaren's upcoming gene-therapy trial and was fascinated. "I really wanted to be involved in gene therapy because it has a great potential to benefit patients," she says. So she took time off from her residency to pursue a PhD in MacLaren's lab. Three and a half years later, she is in the final stages of her thesis project on gene therapy for retinal degeneration. After she completes her residency, she hopes to combine her research with a career in clinical ophthalmology.

These successes have helped to restore investor confidence in the field. "The investment community has started showing up again at scientific conferences for gene therapy," says Gruber. "The attendance went from basically all scientists to about five investors for every scientist."

As a veteran in the field, Gruber has a broad view of its opportunities for researchers. In 1987, he co-founded Viagene, one of the world's first gene-therapy companies, in San Diego. The company had some early successes in treating melanoma and was acquired by Chiron in 1995. But growing concern about gene therapy during the early 2000s and a subsequent acquisition by Novartis in 2006 meant that the research was halted. In 2007, Gruber and several colleagues founded Tocagen.

Dinah Sah says that candidates would ideally have industry experience in adeno-associated viruses.

VOYAGER THERAPEUTICS

He notes that renewed investor interest has spurred the founding of companies and the growth of existing ones like his, leading to a proliferation of jobs. "We have openings for scientists in almost every department in our company, at several educational levels," he says. He also says that gene therapy is a multidisciplinary activity and as such, there are opportunities for people with a variety of degrees, including molecular biologists, cancer biologists, biochemists, pharmacologists, clinical researchers and engineers.

Voyager Therapeutics in Cambridge, Massachusetts, is another company benefiting from the renaissance. Launched in February with the help of US$45 million from Third Rock Ventures in Boston, Massachusetts, the start-up aims to treat disorders of the central nervous system, such as Parkinson's disease and motor neuron disease, by boosting or inhibiting the expression of genes linked to the conditions. The company uses AAVs to deliver the therapeutic genes to the brain.

TOP OF THE HEAP

Unlike other popular vectors, AAVs can target non-dividing cells, such as those in the central nervous system. Their genomes do not readily integrate into the target cell's genome, so the cancer risk is minimal and they are less likely to trigger an immune reaction than other types of virus. "AAVs have been used in more than 1,300 patients in clinical trials, and no serious adverse events have emerged," says Jeff Goater, vice-president of business development at Voyager. "I think it's fair to say that they have risen to the top of the heap in terms of the viral vector of choice for a number of therapies."

Dinah Sah, senior vice-president of neuroscience at Voyager, says that the company is advertising for scientists, specifically those with expertise in working with AAVs as well

as knowledge of biology and diseases of the central nervous system. In addition, skills in molecular biology and protein chemistry are highly valued. "We'd prefer someone with a PhD or someone with a master's or bachelor's degree who has at least seven years of industry experience," she says.

Goater agrees that industry experience is prized but says that most of the candidates with AAV experience tend to come from academia, not industry. As recently as five years ago, he says, gene-therapy companies were focusing on other viruses and much of the work on AAVs was being done at universities. People interested in working in the gene-therapy field should therefore gain experience with this up-and-coming vector, he advises.

Perhaps the biggest boost to the field came in 2012 when the European Medicines Agency approved Glybera (alipogene tiparvovec) for the treatment of lipoprotein lipase (LPL) deficiency in patients with severe or recurring pancreatitis. Glybera uses an AAV to deliver a working copy of the LPL gene to muscle cells. It joins Gendicine, a recombinant adenovirus approved in China in 2003 for head-and-neck squamous-cell carcinoma, as the only gene therapies to obtain regulatory approval thus far.

"The approval of Glybera dramatically changed the landscape in the field of gene therapy," says Jörn Aldag, chief executive of UniQure, Glybera's developer in Amsterdam. "It's the first time that both pharma companies and investors recognized that gene therapy is here to stay."

Aldag hopes to obtain US Food and Drug Administration approval for Glybera by 2017. UniQure also has therapies for diseases such as haemophilia and Parkinson's disease in clinical trials. The company has grown from 45 employees in 2012 to 140 employees, and expects to add 50 more to its workforce by the end of the year, Aldag says.

Although opportunities seem to be on the rise, some are likely to remain cautious — what is to prevent the bubble from bursting again? Levine cites an information-technology concept known as the Gartner Hype Cycle. "Whenever there's a new technology, it goes from a peak of inflated expectations, to a trough of disillusionment, to a slope of enlightenment, and a plateau of productivity," he says. "Gene therapy is now in the enlightenment stage."

Levine says that the current period of growth is different from that in the 1990s because of the accumulated clinical experience. "Thousands of people have now been treated with gene therapy," he says. "And we have much better tools, techniques and equipment than we had back then." ■

Laura Cassiday *is a freelance writer in Hudson, Colorado.*

TURNING POINT
Ashvin Vishwanath

Ashvin Vishwanath, a condensed-matter physicist at the University of California, Berkeley, received a Guggenheim fellowship in April for his exceptional scholarship. It will allow him to spend several months trying to fabricate the exotic states of matter that result from interactions between quantum particles. He describes how reaching out to colleagues in other fields transformed his career.

Why choose a career in physics?
Growing up in India, I realized that it was not common to pursue a pure-science career. I did my master's at the Indian Institute of Technology Kanpur, where 90% of students were engineers. My choice of condensed-matter physics was also unusual — my peers were more attracted to particle physics or string theory. I wanted to be able to conduct experiments to test my theories.

How did you approach your PhD?
I did my PhD at Princeton University in New Jersey on high-temperature superconductors — specifically, how their structure differs from that of regular superconductors. The electrons look like a pair of dumb-bells rather than a pair of circles as in other types of superconductors. My thesis explored the consequences that arise from this pattern. No single one of my PhD papers was spectacularly received, but my colleagues noticed that I was doing a lot of work independently — framing problems and finding solutions on my own as well as working with my adviser and other postdocs. I got a number of postdoc offers.

How did you choose which postdoc to accept?
I made my decision on the basis of potential collaborators, because I felt I would do better science working with someone that I could discuss and generate ideas with. During my PhD, I noticed a paper by Senthil Todadri, a condensed-matter physicist at the Massachusetts Institute of Technology (MIT) in Cambridge. I e-mailed him with some questions, and we launched a collaboration on superconductivity. We got to know each other scientifically and, ultimately, I accepted a postdoc at MIT so that we could continue our work — I had the right instinct for what was important for a longer-term perspective.

What did you work on?
We studied the properties of phase transitions and showed that a seemingly implausible phase transition in superconductors became plausible at the quantum level. The e-mail that led to this collaboration and to this breakthrough finding was therefore a turning point for me.

Did you jump at the opportunity to apply for tenure-track jobs?
No. A few universities encouraged me to do so, but I wanted to spend as long as I could as a postdoc — I didn't think I had a discovery that was significant enough to give me the momentum necessary to start a successful lab. Eventually I landed a job at the University of California, Berkeley, which let me delay my start by a year to get more time as a postdoc. It was during that year that Senthil and I, along with other collaborators, discovered a new phase transition in a magnet, which is a relevant starting point for work in high-temperature superconductors. Had I rushed into a faculty position, I would have missed one of the most productive times of my career. When I'm making career decisions, focusing on the science has always worked best for me. Ten years on, I keep returning to the research questions I asked during my postdoc.

What will you do with the fellowship?
My group proposes the existence of states of matter that have properties that currently exist only in theory. These states obey the laws of nature, but I want to see if we can realize them in a material or synthetic system made of atomic gases. For example, I want to find a system that is a three-dimensional analogue of graphene.

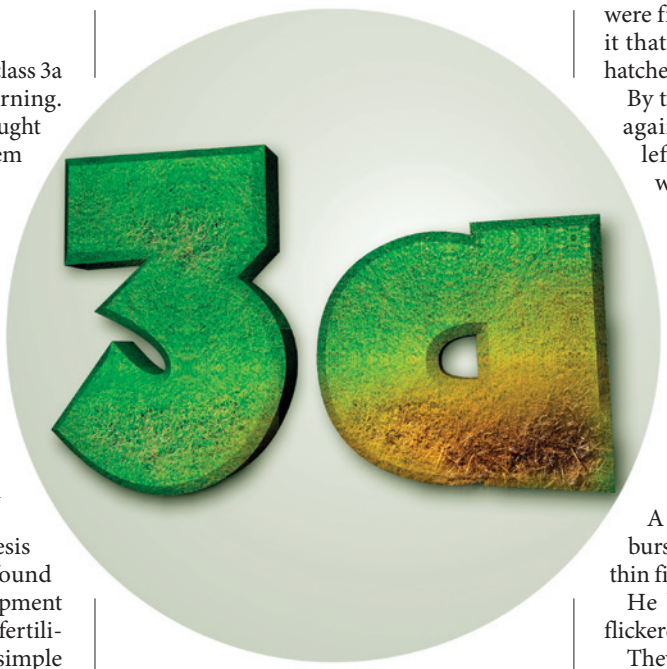
Could this fellowship be a turning point?
Yes. If we can eventually make these materials, it would be huge. That's the dream of every theoretical physicist — to one day bring together a beautiful theory and the experiments to prove it. ■

INTERVIEW BY VIRGINIA GEWIN

VARIATIONS

The importance of a good education.

BY WILLIAM MEIKLE



JACEY

They took Johnny Green from class 3a at ten o'clock on Tuesday morning. He was the last to go. They thought I hadn't noticed, but I'd been onto them for a while now.

Everything went normally, just the usual daily grind in the classroom. Until 3a came round to biology last month. Jack Doyle asked me about cloning, and that got us into a discussion on ethics and Frankenstein foods that actually had most of the class interested for once … apart from Jack, and Mary Brown. She had taken on a quizzical look.

"Sir," she said. "Can you explain parthenogenesis to us?"

"Certainly," I said. "Parthenogenesis is a form of asexual reproduction found in females where growth and development of embryos or seeds occurs without fertilization by males. It happens a lot in simple plants, and has also been shown in some snakes and amphibians."

Mary put her hand up.

"Sir. Is it true that the offspring produced by parthenogenesis almost always are female?"

I nodded.

She put her hand up once more.

And it happened.

Mary Brown flickered. I was looking at a gnarled green thing. Instead of a hand, it waved a shoot above the main body of a squat trunk, a shoot with five thin branches, each tipped with a hard thorny edge.

"I want to learn," she said.

"We want to learn," Jack Doyle added.

"Teach us. Teach us now."

I managed to hold myself together till the end of the class, but by the time I got to the washroom I felt ready to scream. I splashed cold water on my face, and gave myself a long hard stare in the mirror. I didn't look crazy, but it felt as if reality was slowly draining away.

I had 3a again the next day.

By now I knew what I was looking for. Three more had been taken, and had moved to the left side of the room to sit with Jack Doyle and Mary Brown.

I started the day's lesson … one on speciation, but that wasn't enough for them. Their hands shot up almost as soon as I began.

"Tell us about mutation," they said as one.

"How has punctuated equilibrium been a driver for the selection of intelligence?" they asked.

Whatever they ultimately wanted to know, they were learning fast.

I taught them. And while I did so, I learned more about them. By looking out of the corner of my eyes and letting my vision go slightly out of focus I could see them … not clearly, but enough. They were still the same gnarled mass of green plant-like matter. But as they listened to me, small shoots rose and wafted in the air. Tiny nodules on the fleshy trunks swelled and pulsed. Some looked ready to burst.

The other children … the human children, knew something was wrong. They kept looking at me, and back at the small group of what seemed like their classmates.

I wouldn't acknowledge their fears. I couldn't — to do so would give away the fact that I knew them for what they were.

So I kept teaching, and they kept asking questions.

"Tell us about morphic resonance," they said.

And at that they had me stumped. The words had some meaning to me, but I couldn't quite place it. It was near the end of the lesson though, and I was able to talk my way through to the bell.

I could tell they were frustrated, but so was I. I read up on it that night to be ready for them. I also hatched a plan.

By the time 3a came round for biology again, Johnny Green was the last one left. He entered the classroom like a whipped dog, and slunk into a seat near the front.

The rest were so confident that they took him, right in front of me. I watched out of the corner of my eye as I talked, telling them what I'd learned of morphic fields, explaining how it wasn't a recognized scientific paradigm. Even as I spoke, tendrils crept across the room.

Young Johnny never saw it coming. A pustule on one of the green shoots burst, and Johnny's face was covered in a thin film of spores.

He breathed in, coughed once, and flickered.

They had him.

"Tell us about forced mutation," they asked.

"Tell us about genetic manipulation," they demanded.

"We need variation," they said.

I gave them a new word. I taught them about chemomute, a chemical reagent used to bring about targeted genetic mutation.

"Variation," they said. "Teach us."

I showed them a vessel full of a thick liquid.

"Building variation into populations has been something scientists have known how to do for some time," I said. "And it is best demonstrated by an experiment. I think it's time we did some practical work."

Tendrils waved in excitement.

I taught them about the importance of hypothesis, experiment and results gathering.

I taught them how to make chemomute.

I taught them how it would bring about natural, spontaneous beneficial variation in any species that used it.

Then they drank it.

There are two things I didn't teach them.

The first is that even teachers can lie.

The second is that biology teachers know how to make weed killer. ■

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

William Meikle *is a Scottish writer, now living in Canada, with 20 novels published in the genre press and more than 300 short-story credits in 13 countries.*